DISSERTATION

RAND

Hands-on Science and Student Achievement

Allen Ruby

RAND Graduate School

DISSERTATION

RAND

Hands-on Science and Student Achievement

Allen Ruby

RAND Graduate School

The original version of this study was prepared as a dissertation in December 2000 in partial fulfillment of the requirements of the doctoral degree in public policy analysis at the RAND Graduate School. The faculty committee that supervised and approved the dissertation consisted of Brian Stecher (Chair), Dan McCaffrey, James Dertouzos, and Stephen Klein.

The RAND Graduate School dissertation series reproduces dissertations that have been approved by the student's dissertation committee.

Published 2001 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, Virginia 22202-5050
201 North Craig Street, Suite 102, Pittsburgh, Pennsylvania 15213
RAND URL: http://www.rand.org/

To order RAND documents or to obtain additional information, contact RAND Distribution Services (Telephone: 310-451-7002; Fax: 310-451-6915; or E-mail: order@rand.org). Abstracts of all RAND documents may be viewed on the World Wide Web (http://www.rand.org). Publications are distributed to the trade by National Book Network.

Hands-on Science and Student Achievement

By

Allen Ruby

Abstract

From the late 1950s through today, hands-on science has been promoted as a method of science instruction. Currently, recent national science reform efforts seek to temper its role. However, no consensus has been reached on the relationship of hands-on science to student achievement though this topic has been researched since the turn of the 20th century using various methods. To improve upon the literature, this work addresses three major limitations of past research— the lack of data on performance assessments of student achievement, the need to control for factors affecting both hands-on science and test scores, and the potential for a differential relationship by student ability. This work focuses on three research questions: 1) whether hands-on science is positively related to student achievement as measured by standardized test scores using both multiple choice and performance tests, 2) whether this relationship is stronger when using performance tests, and 3) whether this relationship differs by student ability. We apply regression analysis to two data sources. The primary data set is the 1994 RAND Survey of 1400 8th grade students and their teachers in Southern California which includes multiple choice and performance test scores. A second data source is the nationally representative NELS:88 with a focus on the 8th grade student sample. The initial findings vary by source of report, student or teacher, on the level of hands-on science. When accounting for the quality of the reports, the results show an association between the level of hands-on science and student test scores for both multiple choice and performance tests. The results find little difference for this relationship by type of test. Nor do they show strong evidence for a differential relationship due to student ability. These findings support the promotion of hands-on science at the middle school/junior high level while raising a concern about current science reform attempts to reduce and redirect its use. They also provide little evidence to support performance test programs on the grounds that they better reflect what is learned though hands-on instruction. Caveats on the findings and further research needs are discussed.

Table of Contents

Chapter 4 Tables & Figures

Chapter 5 Tables

Chapter 1: Introduction

This dissertation examines the relationship between "hands-on science", a method of science instruction, and student achievement. We use the term "hands-on science" to include all hands-on activities carried out by students during their science class. The public policy debate over hands-on science in the U.S. goes back a century and primarily reflects a changing focus in theories and educational goals, and political influences rather than research findings of how hands-on science is related to student achievement. This dissertation attempts to fill this gap through a systematic examination of the relationship between the hands-on science students experience in the classroom and their science test scores in standardized multiple-choice tests and standardized performance tests.

Hands-on science has been on the rise from the 1970s as can be seen in a comparison of teacher surveys on hands-on science in 1977 and in 1996 shown in Table 1-1 below. The table shows large declines in the response "never or hardly ever", large increases in the responses "1-2 times a month" and "1-2 times a week", and smaller declines in the response "almost every day". Overall, we see about a 30 - 40% rise in the percentage of teachers reporting at least weekly use of hands-on science.

Table 1-1: Teacher Reported Use of Hands-on Science[1]

Frequency of Hands-on Science	Grades 4-6 1977	Grade 4 1996		Grade 7-9 1977	Grade 8 1996
Almost Every Day	12%	9%		24%	18%
1-2 Times a Week	27%	47%		38%	62%
1-2 Times a Month	27%	42%		17%	18%
Never/Hardly Ever	35%	3%		21%	2%

[1] The data from the 1977 survey was collapsed from five to four categories and the percentages were recalculated to account for missing responses which were not reported in the 1996 data. Sources: Weiss 1978; Sullivan and Weiss 1999).

More recently, the centrality of hands-on science to instruction has come under debate for theoretical, practical and political reasons. Theoretically, there are concerns as to whether hands-on science is an optimal method to teach science and whether it best supports the goals of education. Two publicly funded national initiatives, Project 2061 of the American Association for the Advancement of Science and the National Science Education Standards of the National Research Council, are working to reorient the goal of science education toward scientific literacy for all students using an instructional method known as inquiry which includes a reduced role for hands-on science in instruction. Practically, hands-on science faces a challenge due to its relatively large time requirements per topic from concerns that students may learn more topics through other teaching methods. Politically, there are concerns whether hands-on science is of equal benefit to students of differing ability.

Past research has not provided satisfactory conclusions on any of these issues even the more fundamental one of the relationship of hands-on science to student achievement. Overall, small-scale experimental studies have not found a positive link between hands-on science and student test scores. Studies of the impact of curricula with a large hands-on science component show a positive link between curricula and higher test scores but it is not clear whether it is the hands-on component that is responsible for such a link. Findings based on national and international surveys have been inconsistent and inconclusive. Limitations in this work may contribute to the lack of conclusive findings. Sources of these limitations include: (1) a lack of control for factors that are linked to both hands-on science and achievement, (2) a lack of performance tests that may better test the association of hands-on science to students' achievement, and (3) a lack of attention to how different levels of student ability affect the relationship.

A goal of this dissertation is to address these limitations in previous research. Building on the literature, we specify three research questions. The first question asks whether there is a positive overall relationship between hands-on science and standardized test scores for both multiple choice and performance test scores. An overall assessment of the relationship between hands-on science and student achievement would inform educators and policy makers whether hands-on science should be emphasized. We use two types of standardized tests to measure achievement. Multiple-choice tests are widely used. Their broad content coverage helps us consider the concerns over narrowness of content coverage associated with hands-on science. Performance tests though less widely used help us address the concern that certain important skills are taught well using hands-on science but are better tested using performance rather than multiple choice tests. Evidence of a relationship would be stronger if found in both types of tests.

Our second question asks whether the relationship between hands-on science and achievement is stronger for performance tests than for multiple-choice tests. By examining the relative size of the relationship between hands-on science and multiple-choice tests vs. performance tests we can address the question whether performance tests are a better measure of the outcomes of hands-on science. These results will contribute to the current debate over the need for wider use of standardized performance tests.

Third, we ask whether the relationship between hands-on science and achievement differs by student ability. At a minimum, support for broad use of hands-on science requires a positive relationship for at least one group of students and a neutral relationship with others. For example, if hands-on science benefits lower-ability students without hurting higher ability students, a policy emphasizing hands-on science will enhance educational equality without imposing a negative consequence on higher-ability students.

To address these three research questions we use two recent sources of data, the RAND 1994 Survey and the National Education Longitudinal Survey of 1988 (NELS:88) that compliment and supplement one another. The RAND data set surveyed 1400 8th graders and their teachers in Southern California. The strength of this study is that the students took both multiple choice and performance science tests allowing us to compare the relationship of hands-on science to each. NELS:88 surveyed approximately 25,000 students in 8th, 10th and 12th grades and their teachers. The sample is nationally representative to allow generalization to the whole population. Multiple-choice science test scores are available for the 8th, 10th and 12th grades, which allow us to examine the grade difference in the relationship between hands-on science and student achievement. NELS:88 includes broad information on teachers and students, allowing us to control for factors influencing both hands-on science and student achievement for an accurate assessment of the relationship between hands-on science and student achievement.

Our analysis will directly address the limitations of previous research. First, the analysis controls for variables linked to both hands-on science and student achievement, such as race, sex, ability, socioeconomic status, school environment, and course taking. Using a multivariate rather than bivariate method while controlling for these variables, we will be able to estimate the "true" net association of hands-on science and student achievement. Second, exploiting the advantage of the RAND 1994 data that includes both multiple choice and performance test scores for the same students, we develop a statistical method to directly estimate the relative size of the impact of hands-on science on multiple-choice tests vs. performance tests of the same students. Third, we examine the relationship of hands-on science to student achievement for students of different ability levels and in different grades by including interaction terms between hands-on science and ability groups and by comparing the estimates for different grades. In addition, we undertake several sensitivity analyses to tests of the robustness of our findings.

The structure of the dissertation is as follows. Chapter 2 provides background on hands-on science including its formal definition, the continuum of instructional approaches for its use, the history of its promotion, and the current policy debate over its role in science education. Chapter 3 combines a review of the theoretical rationales for hands-on science's link to student achievement with a review of the past research on this topic to develop hypotheses regarding our three research questions. In addition, the two data sets, RAND and NELS:88, used to test the hypotheses are briefly introduced. Chapter 4 presents the analysis that tests the hypotheses using the RAND data set, including measurement, model specification and interpretation of results. Chapter 5 does the same for the NELS:88. Chapter 6 summarizes the major findings from the two data sets, makes concluding remarks, offers policy implications, and discusses opportunities for further research.

Chapter 2: Background on Hands-on Science

In this chapter we provide background information on hands-on science that will help shape our analysis. Specifically, we provide the definition of hands-on science, a typology of the instructional approaches used with it, and a history of its promotion including the current policy debate over its role in science education.

We define hands-on science noting that it does not represent a completely new idea in the literature but broadens the meaning from past terms such as "lab", to encompass a wider range of settings from the lab to the classroom, or "experiment", to include a wider variety of activities that may not be actual experiments, such as observing or measuring.

We offer a typology of the instructional approaches that can be used with hands-on science. These approaches fall along a continuum based on teacher and student roles. That there are many approaches increases the complexity of linking hands-on science and student achievement

A history of the promotion hands-on science shows that we are currently in the midst of a time of debate over its use due to a recent historical trend supporting its use and new initiatives, some of which continue this support while others are attempting to temper that support. From a policy perspective, this makes it a good time to consider the issue as to whether there is empirical evidence supporting either the current rise in support or the reconsideration of its use.

Definition

Traditionally, the terms "laboratory" or "experiment" have been used to describe practical work done by students during science class in place of such other methods of instruction as lecture, reading, recitation, worksheets, teacher demonstration and more recently, computer simulation. These two terms are somewhat limiting for two reasons. First, many students, especially in primary and middle school, do not have access to a laboratory but perform hands-on science activities in their regular classroom. Second, students may carry out hands-on activities that are not actual experiments, for example observation and measurement[2]. The term "hands-on science" includes all such hands-on activities carried out by students be they experiments or not and be they done in the classroom or in a laboratory. The term captures a broader array of student activities we want to investigate and avoids some of the limitations created by the narrow definitions of traditional terms. The term defines a specific method of instruction, based on activities carried out by students, but its use does not preclude other instructional methods for it is often used in conjunction with them. But as class time is limited, the greater the use of hands-on science the less time is available for other methods.

Instructional Approaches for Hands-on Science

Different instructional approaches can be used with hands-on science. Which approach is considered most appropriate has varied over time. And even when one approach dominates, teachers will use the others as well in their classrooms. This

[2] Science modules containing both lesson plans and the hands-on materials to be used by students became available beginning in the 1960s, see more details in the history section, contributing to the use of a broader term "hands-on science." They include many non-experimental activities and have made it easier for teachers to implement hands-on science within a regular classroom.

variation in primacy among and combination of the approaches adds another facet in considering the relationship of hands-on science to student achievement. Here we describe the major instructional approaches for hands-on science, which include: verification/demonstration, discovery, exploration, inquiry, and process skills.

Overall hands-on science has primarily been used to verify or demonstrate a phenomenon in support of direct teaching. Usually the phenomenon is described first in lecture or by the textbook and the students then carry out a well-specified (by the teacher or lab manual) activity that allows them to see the phenomenon or some aspect of it. The other instructional approaches to using hands-on science recognize the usefulness of verification in making an abstract concept concrete and consider it a complementary approach. The verification approach has been criticized on two counts. Its overuse may waste time on repetitive actions. Its recipe manner allows students to only follow directions and watch the results without having to use their own abilities to understand what should happen, how to do it, and what it means (NEA 1920, AAAS 1997).

The discovery approach, in contrast to verification, provides students with materials to work with but little direct guidance on what to do or what is expected to be found. Discovery has two goals. First, it is expected that students will discover phenomena on their own and will understand and remember them better by doing so than if they were shown. Second, the act of discovery will convey how science is carried out in practice (Bruner 1960). Discovery has received both practical and theoretical criticism. In practice, discovery proved too difficult for students to implement on their own. The need for increased guidance by teachers led to a name change in the approach to “guided discovery” (Hodson 1996). Additionally, its applications are limited in that some concepts cannot be discovered by students in school settings (e.g. the atomic theory of matter). Discovery has also been criticized from a philosophy of science perspective in giving a false view of science as using only inductive thinking (Hodson 1996).

The exploratory approach to hands-on science may appear similar to the discovery but is actually more closely linked to the verification. Students are first given materials to handle with little guidance or expectations. The goal is to make them comfortable with the topic, stimulate their interest, and encourage them to raise questions. Following the exploration phase, they receive direct instruction in the topic. Unlike discovery, there is no expectation that the students will discover the underlying concepts, though they may identify issues and questions that can be addressed during direct instruction.

The inquiry approach, like discovery, also contains the two goals of learning specific concepts as well as developing the capacity to carry out inquiries on one's own. The later goal involves teaching the student the set of thinking and doing skills plus their overall use in problem solving while addressing a specific topic, often of student choice to increase student interest. The teacher's role is to provide support and guidance especially through questioning rather than leadership though this fluctuates with the ability of the students. Inquiry differs from discovery in that it recognizes the use of both inductive and deductive methods. It differs from the process skills approach (described below) by its attempt to teach an overall method that incorporates process skills rather than addressing them separately. Students do not have to discover all knowledge on their own. Hands-on science is just one technique that can be used in inquiry. Unlike the verification approach, inquiry hands-on science activities are not recipe in nature, the outcome is not to be known ahead of time, and the student is to have an active role in designing, carrying them out and interpreting the results. In practice, inquiry is a difficult and time-consuming approach which demands great skill and knowledge from the teacher and is difficult to package in a standardized curriculum.

The process skills approach attempts to teach individual processes used in science without regard to any specific science topic or discipline. Hands-on science is the primary technique used in teaching those processes that require hands-on activity, e.g. measurement. The process skills approach came under attack in later years based on arguments that content was non-excludable from process because of the need for content in problem solving, because of the difficulties in transferring process skills from one context to another, and because students did not appear able to assemble the individually taught skills into an overall ability to problem-solve (Champagne, Klopfer, Gunstone 1982; Hodson 1996). Teaching process skills remains a goal of science education and they continue to be taught both separately and in conjunction with content matter. A recent thread in this approach is the proposal to teach the history and philosophy of science as a way of instructing students in the overall process used by scientists and to show that the processes used may vary by the type of field and scientist (Matthews 1994).

In sum, there are different instructional approaches to using hands-on science and a teacher may use any combination of them over the course of the year. Historically, the emphasis on each has varied. Today, inquiry is the primary approach proposed but we continue to find the other approaches in use both in the classroom and in the new curricula aligned with the current efforts at science education reform (see History section below). In Chapter 3 we further discuss how the variation in instructional approaches affects the research on the relationship of hands-on science to student achievement.

History of the Promotion of Hands-on Science

We are currently in the midst of a time of rising support for the use of hands-on science as well as a time of reform in science education, a reform that is reconsidering of the role of hands-on science. It is an opportune moment to consider whether the assumption of the positive relationship of hands-on science to student achievement holds and should continue to be given great weight as reform occurs.

The level of promotion of hands-on science has swung widely over time due to varying causes. The arguments behind hands-on science cycle through different periods: past arguments are ignored then resurrected. Similarly, the instructional approaches to using hands-on science have varied in their acceptance. The debate over how to use hands-on science is often hidden by the issue of whether to use it, but swings within the former are just as wide as those of the latter. This section documents the historical changes in both the promotion of hands-on science and the instructional approaches to its use. This history focuses on attempts to promote the use of hands-on science and it is important to remember that throughout the period covered, commentators have noted the primacy of textbooks, lecture and recitation and the lack of hands-on science in the classroom (Deboer 1991).

The issue of hands-on science in the classroom is not a new one. The debate over it can be traced back to the struggle to introduce science into the primary and secondary curriculum that took place during 1800s. At the beginning of the 19th century, primary school (which then included grades 1-8) focused on reading, writing and arithmetic and secondary school focused on Greek and Latin. Supporters of this classical curriculum argued that generalized mental exercises (such as memorizing words or doing proofs) increased mental capacity while providing a humanistic and refined education.

Proponents of science in the curriculum tried to both expand upon and supplant this reasoning with three arguments. First, learning science was argued to be a better form of mental exercise for it required a wider range of mental abilities to not only memorize facts but to organize them into generalizations and use them in inductive thinking. Second, at this time there developed a new view of learning which saw it as a process by which neural connections were built and strengthened by organizing sense perceptions and building generalizations from them. To succeed, learners had to receive information in an organized manner plus have it repeated in different contexts and combinations. Successful education then required both effort and organization, and science education would provide a means for this organization. Third, science education was proposed as useful in everyday life. On one hand, it would provide information directly relevant to the student's life and work, such as information on maintaining good health or increasing crop production. In addition, it would provide a system of thought, using inquiry to discover facts, that the student would use throughout their life independent of a teacher. These arguments proved successful and by the end of the 19th century the question switched from whether or not to include science in the curriculum to how much science and which disciplines to include (Deboer 1991).

The proponents of science education considered the science laboratory central to their view of education in contrast to the predominant use of memorization and recitation from textbooks. Almost all the arguments used today in support of student hands-on science were first offered in the 1800s. These arguments touched upon children's abilities, improved understanding, skills and long-term learning. Young children were thought to easily accumulate facts and the simple relationships between them through sense impressions. Students would better understand the meaning of written words through definite images of the phenomena. Reasoning and learning skills, such as making inferences and judgements and verifying them, would be learned through observation and experiment. Through such student investigations, personal understanding would occur rather than only memorization, the material would be retained longer, and students would become independent learners. (Huxley 1899, Spencer 1864, Youmans 1867). Formalized approaches to using hands-on science followed, for example the techniques of Johann Pestalozzi, which focussed on the study of natural objects, and those of Johann Herbart, which included a first step of pupil experience with the natural world (Deboer 1991).

While the proponents of science education in the 19th century agreed on the centrality of the science laboratory they did not agree on how hands-on science should be used. Rather they formed two continuums over its use and these remain today. The first continuum considers the instructional approach for using hands-on science and was clearly set out by Smith & Hall in 1902 though additions were made over time. At one end is an instructional approach called discovery in which through laboratory activities students would discover both facts and concepts individually and independently of the teacher so as to become independent learners. At the other end is the teacher or lab manual-directed use of hands-on science to verify principles, to illustrate them and make them more vivid in the minds of students. In between are two other approaches. Under the exploratory approach, students first explore a topic through hands-on activities then return to teacher directed instruction. Second, through an inquiry approach students answer questions by generating their own investigations with the assistance of the teacher who guides them through questioning.

The second continuum concerns the goals of science education: what type of science knowledge is to be taught and, as a result, what uses can it be put to. At one end is science as a structured body of knowledge, logically organized and primarily concerned with facts and the underlying principles and theories. The goal is to give the student a firm grounding in the basics of science. At the other end is a very applied education linking science either with meeting individual needs (e.g., understanding proper nutrition or providing employment-related technical applications) or addressing societal issues, such as pollution. The applied approach is also concerned with practicality as a way of increasing student interest in science by making it relevant to daily life. In between is the goal of teaching students how to use the skills and techniques of science as an investigative process. These skills can help a student contribute to the body of scientific knowledge and form a tool the student can use in daily life, be it for personal decisions, at work or in an attempt to solve societal problems.

While there was not full agreement along either continuum, early proponents of science education proposed that students spend much of their time in the laboratory. For example, in 1892 a committee of ten college and school leaders was appointed by the National Educational Association to determine college entrance requirements including the best method of instruction in both high school and college. The committee established conferences in each subject area. The Conference on Physics, Chemistry and Astronomy determined that the majority of elementary school science should be done through experiments and that 50% of the work done in high school should be carried out in the lab. The Conference on Natural History determined that 60% of the time should be in the laboratory and the entire course should focus on the observations made in the lab (Deboer 1991). Not all science proponents supported the idea that science education should take place primarily in the lab. For example, Edwin Hall (Smith & Hall 1902) while a firm supporter of teaching through the laboratory, argued that it contained too many flaws to be used alone. He saw learning occurring too slowly in the lab and students focusing on the mechanical part of the experiments while ignoring the hard thinking that had to be done to understand the point of them. For him, the laboratory was only one technique that should be combined with lectures, recitation, demonstrations and numerical problem solving.

The rise of progressive education (1920s – 1950s) led to a shift in important ways along the second continuum and a decline in the importance of the laboratory. The goal of developing an individual's intellectual skills shifted to the goal of developing individuals who would contribute to society. Science was to have direct consequences for everyday life (NEA 1920). The switch to a focus on applications to everyday life was also based on changing instruction to meet student interests to increase student motivation and to connect new knowledge to what the student already knew. Previously, everyday applications had been considered useful for increasing student motivation to learn and illustrating concepts but now they became the focus of science education.

The change in focus affected science education in several ways. Teaching through themes, problem solving and technical applications was promoted with less attention to basic principles. The idea was to organize the material in a way interesting and understandable to the student rather than as the discipline organized it. During this period, the general science course was created in part to provide wide exposure to students who were not completing high school, in part as a way to try to attract more students to specialized courses, and also as a way to implement the new approach to teaching. The disciplines, represented by subject oriented courses such as biology or chemistry, were resistant to change and proponents of the new approach saw the general science course as a way around them. In addition, this period saw the creation of the junior high school and general science courses were extended into 7th and 8th grades (Deboer 1991).[3]

While the new approach also favored direct experience with natural phenomena, the laboratory no longer retained the central role in instruction. Instead, it became one of a number of techniques along with demonstration, lecture, field trips, projects and problems. Laboratory activities were criticized as time wasting, repetitive work and time could be better spent on developing ideas or posing problems or questions that students would be interested in answering (National Society for the Study of Education 1932). Teacher demonstrations were promoted as a more effective method for providing real world examples (Black 1930; Carpenter 1925) and were adopted by school

[3] One further development during this period that would affect the use of hands-on science was the introduction of the standardized test. Welcomed in part for its ability to differentiate students and its expected ability to determine best teaching methods (such as teacher demonstration versus student lab), a major effect was to focus attention on content mastery which was easier to measure. From that time to today, proponents of hands-on science have argued that such content-oriented tests fail to measure the benefits of hands-on science for student understanding of concepts and their application.

administrations seeking to reduce expenditures. As a result, the use of student labs declined over this period (NSSE 1932).

After World War II, a combination of declining enrollments in secondary science courses plus the competition with the Soviet Union led to a bifurcation of efforts. Greater emphasis was placed on programs to attract talented students into the field of science through discipline oriented courses (PSRD 1947; U.S. Office of Education 1953). At the same time, there were further efforts to revise science education toward the more applied coursework for the majority of students in both secondary school and college who were not expected to work in any scientific field. The latter revisions attempted to also include the discipline oriented courses, mainly chemistry and physics, in secondary school to make the courses more relevant to life situations (NSSE 1947). Additionally, new general science courses were created with such titles as "Consumer Science", "Fused Physical Science", and "Survey Science" (NSSE 1947). A series of courses known as "Life Adjustment Education" was developed for the purpose of replacing traditional subjects such as science and English with courses on life skills (U.S. Office of Education 1951).

The reduced or non-academic focus of the applied and life skills courses led to criticisms and a renewed involvement of scientists in the development of science courses. For example, in 1956 a group of scientists at MIT formed the Physical Science Study Committee (PSSC) under a small NSF grant and began to develop a high school science course (Deboer 1991). The launching of Sputnik by the Soviet Union led to a major federal role in the development of new science courses. Under the 1957 National Defense Education Act, about $700 million dollars was provided to improve science education from 1958-1975. NSF was made the lead agency and it turned to scientists at universities or professional societies for the development of new science courses (Matthews 1994).

A host of new curriculum was developed over this period. In physics, the PSSC published its text and lab manual in 1960, today it is in its 7th edition, and later developed a junior high course called Introductory Physical Science published in 1967. In biology, the American Institute of Biological Sciences organized the Biological Sciences Curriculum Study (BSCS) which published a series of texts in the mid-1960s and continues to operate today. In chemistry, the American Chemical Society developed two courses: 1) the Chemical Bond Approach, and 2) CHEM study aimed at a wider range of students. For earth science, the American Geological Institute established the Earth Science Curriculum Project, published in 1967, and Princeton University's Secondary School Science Project published Time, Space and Matter in 1966 aimed at 9th grade. Three curricula were also developed for elementary school: 1) Science – A Process Approach (SAPA) was developed under the American Association for the Advancement of Science in 1967, 2) Elementary Science Study was developed by the Educational Development Center in 1969, and 3) Science Curriculum Improvement Study published in 1970 and continued today by the Lawrence Hall of Science at the University of California at Berkeley.

In reaction to past curricula, these new curricula contained much fewer technical applications and many were discipline structured. All included a large student hands-on science component. The physics, biology and chemistry curriculums all contained separate lab manuals. The earth science and elementary curriculums were centered around student activities. In some cases, the curriculums came with the materials necessary for the activities thereby addressing the lack of a lab or materials in a school.

These materials did have a widespread effect (though not an overwhelming one) on science education. A 1977 survey of school districts found that about 50% were using new biology materials and less than a quarter were using the new physics and chemistry materials. Overall, about 60% of districts were using one or more of the new curricula for students in grades 7-12 and about one-third of the districts were using new elementary curricula (Weiss 1978).

There was no common instructional approach to how hands-on science was to be done in these new curricula. The same differences in approaches identified at the turn of the century were apparent. The Chem Study course lab program and the Elementary Science Study were based on the discovery approach (Merrill & Ridgeway 1969). Jerome Bruner, chairing a1959 NAS conference on new developments in science and math teaching, had given his support to the discovery approach seeing it as a way to learn the discipline the same way scientists learned it (Bruner 1960). The BSCS was geared to the inquiry approach in part because of Joseph Schwab who played a major role in its development and argued that scientific knowledge changed over time through inquiry and students would better view science in this way if they practiced inquiry as well (Schwab 1962). SAPA was a wholly process oriented curriculum which focused on techniques used in science (e.g. observation, measurement) rather than any specific discipline or facts. SCIS was based on a technique known as the learning cycle which began with the exploratory use of hands-on science to raise student questions leading to direct teaching of the concepts students had experienced and ending with student application of the concepts to other situations (Karpus 1977).

The new curricula also did not solve the disagreement over how much time should be spent on hands-on activities. Shulman and Tamir (1973) list the many different opinions voiced on this topic and note that proposals ranged from half of class time to not very often.

The new curricula represented a greater federal role in science education increasing the importance of the issue of hands-on science as a national public policy issue. While much of the previous debate and action over hands-on science had occurred within the education community and at the state, district and school level, the federal government now had taken a major role in supporting the development and promotion of curricula containing a significant hands-on component for national use.

By the mid-1970s there was a turn away from these nationally supported curricula. In part politically-motivated, in part based on the failure to increase enrollment in science courses (Deboer 1991), and in part with the realization that developing curricula does not ensure they are correctly implemented especially if logistical support and adequate training of teachers are not also provided (Arons 1983). In response to the discipline-orientation of many of the curricula, there was a resurgence toward making science relevant to students' lives and focusing on socially-relevant issues such as the environment. A scientific literate student who could use science concepts, process skills, understood that value judgements were made in science and understood the links between science, technology and society was proposed as the goal of science education. Students should be able to understand the daily world and have the tools to learn more (NSTA

1971, Hurd 1970). One branch of the science literacy approach, known as Science-Technology-Society, called for the organization of the science curriculum around social issues (Hofstein & Yager 1982).

The scientific literacy approach become embodied in two national efforts to reform science education both supported by the federal government. These efforts included the further step of addressing what content should be taught and how to teach it.

In 1985, the American Association for the Advancement of Science began a long-term effort known as Project 2061 to turn U.S. education toward a scientific literacy approach. For AAAS, scientific literacy encompasses: important scientific facts, concepts, theories, scientific habits of mind, the nature of science, connections to math and technology, the impact of science on individuals and its role in society (AAAS 1997). In 1989, AAAS published Science for All Americans which set out its view of the information a scientifically literate person should know and in 1993 it published Benchmarks for Scientific Literacy which lists specific knowledge and skills to be learned in sets of grades (e.g. k-2, 3-5, 6-8, 9-12)[4] from kindergarten to high school in order to create science literate students. Project 2061 promotes teaching through the inquiry approach. Hands-on science is to make up a minority of the teaching methods used and to be employed possibly no more than once a week (AAAS 1997 chapter 1B). When used, hands-on science is to be geared primarily to the inquiry approach in order to teach scientific inquiry (AAAS 1997 chapter 1B) with some additional use as a method to provide concrete examples of phenomena (especially for younger students) and to practice use of tools (especially for measurement) (AAAS 1989, chapters 12 & 13).

[4] During the 1980s, the middle school replaced the junior high in many parts of the country. Middle schools included grades 5-8 or 6-8 rather than the junior high school's 7-8 or 7-9.

In a separate project, the National Research Council began developing its own set of science education standards for pre-college education in 1991 and in 1996 released the National Science Education Standards. Like the Benchmarks, these standards include topics to be covered within certain grade levels (K-4, 5-8, and 9-12) plus somewhat broader standards on teaching, professional development, program and system standards. Like the Benchmarks, the Standards promote scientific literacy and the use of hands-on science under an inquiry approach (NRC 1996, chapter 3).

Neither the NRC nor Project 2061 are involved in the actual development of curricula nor have they yet set out a clear practical approach to teaching through inquiry (including how to teach hands-on science) that can easily be implanted in a new curriculum. Project 2061 has moved further along this pathway and has begun to review science curricula, starting with commercial textbooks using a set of 23 criteria, one of which has a clear link to hands-on science[5]. Over the next several years, it is expected that the current reform efforts will refine and consolidate their approaches including the role of hands-on science and the suggested approaches for its use.

The late 1980s also saw the return of NSF as a major actor in the creation of new science curricula and promotion of their adoption. Going beyond its approach of the 1960s in which curricula focused on 1 or several grades, NSF supported the development of curricula that would encompass elementary through middle school (grades K-8). As its first step, NSF funded three organizations to develop activity-based, primary school (K-6) science curriculum. These included: 1) the National Science Resources Center (a

[5] The criteria falls under the Category "Engaging Students with Relevant Phenomena" and reads: Providing vivid experiences: Does the material include activities that provide firsthand experiences with phenomena when practical or provide students with a vicarious sense of the phenomena when not practical? (AAAS 1999).

program founded by the Smithsonian Institution and the National Academy of Sciences), 2) the Lawrence Hall of Science at the University of California-Berkeley, and 3) the Education Development Corporation of Newton, Massachusetts. These three separate organizations have developed a series of science modules centered around student hands-on work and their adoption is expected to reduce lecture and the use of textbooks (NSRC 1997). These grade-level modules cover specific science topics using a hands-on science approach (none of the modules includes a textbook though two have now developed supplemental readings to go along with the activities). The modules come with the majority of materials necessary to do the activities (like some of those developed in the 1960s) thereby making a laboratory unnecessary. They are now sold through commercial publishers. NSF's second step has been to provide additional funding to all three organizations to develop modules for grades 7 and 8 and these are undergoing pilot testing at this time.

Interestingly, in contrast to the AAAS and NRC's proposed use of the inquiry approach with reduced emphasis on hands-on science, the new curricula are heavily hands-on oriented and include a combination of directive, exploratory and process skill[6] approaches along with the inquiry approach. The cause of this divide is partly due to the practical nature of curricula development (it is easier to make a module including materials and activities that are directive or exploratory), partly due to the as of yet failure to operationalize the inquiry approach to the point that it can be packaged in a curriculum used by large numbers of teachers, and partly due to the high level of importance hands-on science retains in the minds of developers and teachers.

In order to help implement the science standards and new curricula, NSF began in 1990 to establish several programs to foster reform of science education at the state, city and district level that include support for the adoption of the new curricula, professional development in their use and other teaching practices that are assumed to support the goals of the National Science Education Standards. These programs include the Statewide Systemic Initiative (grants to 25 states and Puerto Rico), the Urban Systemic Initiative (grants to 20 cities with many children living in poverty), and the newer Rural Systemic Initiative, and together have received over $100 million in funding from NSF (Mervis 1998; Williams 1998). These initiatives foster the adoption of the Standards and new curricula in part by supporting teacher professional development in their use and funding their purchase.

Currently, then, the federal role in science and in regards to the promotion of hands-on science has expanded though not in a coordinated fashion. On one hand, there is a continuation of federal support for the development of new curricula, covering a broader range of grades than before, having a predominant hands-on science focus often provided in a non-inquiry mode. On the other hand, there is now significant federal support for developing and implementing a strategy for the reform of science education including specific attention to content and teaching methods. Along the continuum of instructional approach, current science education policy is in the middle with the theoretical focus on inquiry. Along the continuum of goals of science education science education policy has tilted more toward daily life with the emphasis on scientific literacy. The combination of inquiry and scientific literacy has leads to a view of hands-on science

[6] For example, the new NSF-supported modules developed by the Lawrence Hall of Science address a set of process skills in every content module as well has having two modules that focus specifically on two

as one of many tools that can be used in support of teaching inquiry and scientific literacy. Therefore, attention to promoting hands-on science is to be both reduced and redirected towards making it fit into the new strategy for science reform.

Through the expanded federal role in science education, hands-on science has become a national public policy issue both indirectly, through federal support of the development of new curricula having a hands-on focus, and directly through federal support of efforts to establish national content and teaching standards that de-emphasize the role of hands-on science. Determining the type of association between hands-on science and student achievement would contribute to the resolution of the current conflict between the present theoretical and applied forms of current science reform and fostering a more united federal role.

process skills, measurement and variables, in a content free manner.

Summary

In this chapter we have chosen the term hands-on science to reflect the current trend in public science education towards student activities performed in the classroom that may or may not be full experiments. A number of different instructional approaches can be used with hands-on science. These have risen and fallen in favor over time but a teacher may use any or all in combination in the classroom. Similarly, the promotion of hands-on science has varied in strength over time. The most recent trend has been a rise in the promotion of hands-on science which has been continued in the development of new curricula but is also being tempered by current science reform. Due to the differences in the current view of hands-on science and their implications for science reform, it is an opportune time to consider the link between hands-on science and student achievement.

Chapter 3: Hands-on Science and Student Achievement

In this chapter we examine the theoretical relationship of hands-on science and student achievement and the relationships that have been identified or tested by past research. Based on these, we propose the hypotheses linked to our three research questions. First, we examine the theoretical relationship of hands-on science and student achievement identifying the rationales that have been made for its proposed benefits. Second, we review the empirical literature on the testing of these rationales. As part of this review, we describe several issues that have affected the value of past research and discuss how this analysis will address them. Based on our reviews of the theoretical and empirical literature, we set out the hypotheses to be tested by this work. Last, we provide an overview of the two data sources to be used in testing these hypotheses leaving a more detailed description of each to Chapters 4 and 5.

Theoretical Rationales for Hands-On Science's Role in Student Achievement

Hands-on science has been proposed as a means to increase student achievement in science education. A set of theories has been proposed to explain how hands-on science benefits student learning of science[7]. Science educators identify two broad domains of scientific knowledge: content knowledge and process skills (Glynn and Duit 1995 chapter 1; Lawson 1995 chapter 3). Content knowledge (sometimes called declarative knowledge) includes the facts, principles, conceptual models, theories and laws which students are expected to understand and remember. Process skills (sometimes called procedural knowledge) are the techniques used in science, for

[7] This discussion focuses on learning scientific knowledge and skills. Hands-on science has been justified on other grounds, such as improving lab techniques and motivation to learn (Shulman and Tamir 1973).

example, observation, measurement, and developing hypotheses, which students are to master. Both domains are considered necessary in order for students to fully understand science and be able to apply it (Glynn and Duit 1995; Champagne, Klopfer and Gunstone 1982; Eylon and Linn 1988). Hands-on science has been proposed as a means to increase students' understanding of both types of knowledge.

Scientific content knowledge is often abstract and complex. Examining and manipulating objects may make this abstract knowledge more concrete and clearer. Through hands-on science students are able to see real-life illustrations of the knowledge and observe the effects of changes in different variables. These illustrations also provide references for discussion (Shulman and Tamir 1973; Friedlander and Tamir 1990).

The idea that hands-on science supports understanding of content knowledge is consistent with developmental theory's positing successive stages (from three to five) of mental development through which humans pass. The highest stage includes the ability to work with abstractions. Before this stage can be reached, humans first pass through a stage in which thinking is confined to concrete matters. Interactions with the physical environment (along with other factors) support the mind's passage through these stages (Piaget 1973, Gage and Berliner 1994, Lawson 1995). Under this view, hands-on science can help students move from the second highest stage to the highest stage as it offers concrete illustrations of abstract ideas at a time when the mind needs concrete representations for understanding. Once at the highest stage, however, hands-on science is of much less importance in helping the student gain understanding as the student is now capable of grasping and manipulating abstract ideas.

This argument for hands-on science is also consistent with cognitive theory's information processing model of the mind which includes a long-term memory to store knowledge and a short-term memory to hold knowledge in immediate use. The ability to retrieve relevant knowledge from the long-term memory for use in the short-term memory is based on how the knowledge has been organized in the long-term memory and how strong the associations have been made between individual pieces of knowledge. Hands-on activities create additional associations between pieces of knowledge so that information can be referenced both by its abstract meaning and by a physical illustration of it. In this way, it improves information retrieval (Gage and Berliner 1984).

Hands-on science may also be used to address faults in information processing. According to cognitive theory, the separate bits of knowledge held in the long-term memory are organized using broader concepts known as schema. These schema are organizing principles which guide an individual's understanding of the separate pieces of information and are used to organize and integrate new information. One can form schema that do not correspond to the real world. These misconceptions may prevent learning as new information may be synthesized in a way that justifies the misconception or may be ignored if it contradicts the misconception (Eylon and Linn 1988; Champagne, Klopfer and Gunstone 1982). One approach to instruction, known as conceptual change, attempts to identify these misconceptions, have the student realize they do not accurately explain phenomena and help the student adopt more realistic conceptions. Hands-on science, with its focus on real world phenomena, has been proposed as a method to help a teacher identify these misconceptions as well as provide a setting for students to explore how their misconceptions falsely predict phenomena in preparation for reconsidering them (Driver 1981; Driver and Bell 1986; Friedler and Tamir 1990).

Science education also entails the use of process skills which are the techniques of science such as observation and measurement. What these process skills are and how to teach them affects the use of hands-on science. Currently, a debate continues over many facets of these skills: 1) their number and type, 2) at what age they should be taught, 3) whether they need to be taught in a specific order or at the same time, and 4) whether they can be taught separately from content knowledge. Gagne (1965) identifies eleven skills and places them in two categories, basic and integrated. For him, basic skills must be taught before integrated ones. Lowery (1992) identifies seven skills and attempts to determine at what age students are developmentally ready to learn each. Resnick (1987) argues that higher order skills are used along with basic skills when young students learn and therefore the two types of skills must be taught together in all content areas. In the 1960s a new curriculum, known as SAPA, focused on teaching a hierarchical set of skills without links to content. In contrast, studies comparing experts and novices have concluded that content knowledge is critical to the correct use of process skills (Champagne, Klopfer and Gunstone 1982; Eylon and Linn 1988).

Another concern regarding process skills is promoting the ability of students to use all the individual process skills in combination for problem solving and carrying out a scientific investigation on one's own. Sometimes this goal has been set for only the best students expected to go into the field of science and other times it has been geared to all students through applications in daily life, be they personal or work related. The science education instructional approaches seeking this goal have all included some level of hands-on science.

Some of these process skills are by their nature hands-on, for example measurement, and therefore are considered best learned through hands-on science. Others may be linked to hands-on activities but themselves are not, for example inferring is based on results derived from hands-on activity. Students can learn and practice these skills using results drawn from other sources than in-class activities leading to debate over the need for hands-on science when teaching them. Klopfer (1990), for example, argues for a larger role for hands-on science in learning the skills of gathering scientific information (e.g. through observation and measurement) than in the ability to make inferences and draw conclusions from experimentation.

The theoretical rationales given for the impact of hands-on science on student achievement have not gone unquestioned. Critics argue that hands-on science may reduce student achievement as well as improve it. Whereas proponents argue that hands-on science helps students visualize abstract ideas, opponents argue that it has the ability to confuse as well as clarify. Hands-on science also offers students additional opportunities not to learn as they may be busy doing activities but not thinking about the topic. Additionally, some research has shown that students may not link hands-on activities to written activities concerning the topic being studied (Wellington 1998, Hodson 1996; Atkinson 1990; Resnick and Klopfer 1989).

A practical criticism concerns the time and monetary costs of hands-on science. From Smith & Hall in 1902 through today, critics have argued that hands-on science takes up too much time, requires expensive recurrent purchases, and drastically reduces the amount of material that can be covered in a course.

This practical point raises an issue of equity between higher and lower ability students. Lower ability students are expected to benefit both from the concrete examples and the greater time per topic provided by hands-on science. Higher ability students may be able to understand a topic in a shorter period using less time consuming instructional methods. For them, the additional time spent on one topic when using hands-on science may lead to a reduction in topics covered during a course.

Research on the Relationship of Hands-on Science and Student Achievement

The importance of resolving the question of hands-on science's relationship with student achievement has not gone unnoticed. There is a body of research on the topic, which we review in this section. We organize our review under three broad headings. First, past research has not led to a firm conclusion regarding the link between hands-on science and student achievement. We review the past research with an emphasis on its overall inconclusive nature. Second, we discuss some of the important issues in this research and how the inability to resolve these may contribute to the lack of agreement. Third, we discuss how we plan on improving upon the past work and note that our work will still be subject to some of the same difficulties.

Inconclusiveness of Past Research

Research on the relationship of hands-on science and student test scores has been conducted since the turn of the century. The research has been based on three techniques.

Under small-scale experiments, very small groups of teachers are assigned a particular teaching method (for example, lecture, text based or hands-on science) to use in their classes and the test scores of their students are compared. A second approach is to compare classrooms using different types of curricula. Test scores from students using different curricula are compared and conclusions are drawn based on different levels of hands-on science in each curriculum. Alternatively, surveys have been used to determine the quantity of hands-on science in a classroom and to collect student test scores. The relationships between test score and hands-on science are then examined. Because our work follows this approach, we devote the greater part of our review to this vein of past research.

Overall, reviews of the experimental studies have not found a positive correlation between hands-on science and test scores except for tests of lab skills. Cunningham's (1946) review argued that the benefit of students receiving more lab instruction was in their ability to use lab apparatus but not in their achievement on tests. Shulman and Tamir (1973) reviewed the literature and also found that the majority of studies failed to show lab teaching as more effective than other instructional methods in regards to test scores. The Hofstein and Lunneta (1982) review again found that lab work showed no significant benefits over other methods of instruction when comparing scores on tests for achievement, critical thinking and understanding the processes of science. They did find a positive effect for lab skills. White and Tisher (1985) reviewed teacher and student surveys and found little agreement on the perceived benefits of hands-on science.

Reviews of the impacts of the curriculum developed from the late 1950s through the early 1970s, in which hands-on activities played an integral part, showed more positive results. Bredderman (1983) and Shymansky, Kyle and Alport (1983) carried out meta-analyses to synthesize studies (57 and 105 respectively) on these curricula. The former article focused on elementary programs while the latter addressed K-12 curricula. They reported significant positive results regarding student test scores (both for content and analytical skills). Interestingly, Bredderman (1983) did not find wide variation in the effects of the three elementary science programs: SAPA (process approach), ESS (discovery approach) and SCIS (exploratory approach) but did find some sensitivity to the instructional approach used. For example, students using SAPA scored higher on process outcomes and students using SCIS scored higher on content outcomes. In addition, he found that student outcomes under these programs were higher than those of students using textbooks and only somewhat higher than students using other activity based programs. From these two findings, he concluded that the common features of the three programs, that they contained more hands-on science and gave more attention to process, were more important to student outcomes than any unique features of each program. Initial analyses of the new curricula developed in the 1990s have begun. As the 5th and 6th grade portions of the curricula were developed later and the middle grades curricula are being piloted now, it will be several years before similar meta-analyses can be done to determine the overall and specific effects of these curriculum on student achievement. If the same results are obtained, that will be a strong argument that hands-on science, a common component of the new curricula, is responsible for their benefits.

Much of the recent work on this issue has been carried out using data from international and national science studies. Two organizations run separate series of international studies: The International Association for the Evaluation of Educational

Achievement (IEA) has supported four studies[8] and the International Assessment of Educational Progress (IAEP) has organized two studies. All six of these surveys are similar in that they are cross-sectional in nature, gather survey data from students, teachers and schools, and include student test scores, primarily on multiple choice tests. The results from these studies are mixed and the majority do not show a positive relationship between hands-on science and test scores. In several cases, the value of these studies is hampered by a lack of reporting of the full results.

The IEA's First International Science Study surveyed 10 and 14 year olds and students completing secondary school in 1970-1971. However, only 10 year olds were asked whether or not they made observations and did experiments in science class. In 12 of the 15 countries or regions included in the survey, a yes answer was positively correlated with a higher test score. In the U.S. the correlation coefficient was .18 using a sample of about 5400 students. A further analysis using an OLS regression model with covariates such as student type, school type, home and attitude, found positive coefficients for the variable representing observations and experiments but the significance levels were not reported (Comber and Keeves 1973).

The IEA's Second International Science Study surveyed 10 and 14 year olds in six countries from 1983-1986. The students were asked how often they did experiments in science class and could choose from three answers (never, sometimes and often). Teachers were asked how much time was spent on practical activities, defined as experiments or field work, and could choose from four answers (zero or little, one-fourth, one-half, or three-fourths or more). The teacher item was not included in the U.S. survey.

[8] The fourth study, the Third International Math and Science Study, has not yet released results regarding classroom instructional practices and their relationship with test scores.

The study reported only the average correlation between the hands-on item and test scores for all the countries as a group. For grade 5, the average correlation was .07 using the student responses and 0 using the teacher responses. For grade 9, the test was subdivided into content and process sections. For student responses, the average correlation was .09 for content score and .07 for process score. For teacher responses (with no U.S. participation) the average correlation was respectively .21 and .25 but these were skewed by upward by one country) (Tamir & Doran nd; Doran & Tamir 1992).

From 1983-1987, the IEA carried out its Classroom Environmental Study which included an examination in six countries (not including the U.S.) of how science classroom factors affected achievement scores. Classroom observation was used to collect data on the amount of lab work done. No country showed a positive significant correlation between the amount of lab work and gains from a pre-test to a post-test (five countries had non-significant results and one had a negative significant correlation) (Anderson, Ryan and Shapiro 1989).

The IAEP surveyed 24,000 13 year olds in 5 countries and 4 Canadian provinces in 1988. Students were asked how often they did experiments on their own and with other students. They choose from 5 responses: never, less than once a week, once a week, several times a week, and almost every day. Rather than report the actual correlations of these responses with test scores, the authors stated that the frequency of experiments was not consistently related to test performance (Lapointe, Mead and Phillips 1989, p. 41).

In 1991, the IAEP survey 25,000 9 year olds in 14 countries and almost 52,000 12 year olds in 20 countries. The students were asked the same two questions concerning experiments as in 1988 and their answer was combined into a single measure. No country showed a positive significant correlation between the 9 year olds' responses and their test scores (7 showed non-significant correlations including the U.S. and 7 showed significant negative correlations). Only one country showed a positive significant correlation between the 12 year olds' responses and their test scores (8 had no significant correlation and 11 had significant negative correlations, including the U.S.). The actual correlations were not provided in the study (Lapointe, Askew, Mead 1992, pages 50 and 94).

Domestically, the U.S. has two national surveys that include science education. The National Assessment of Educational Progress (NAEP) is an ongoing cross-sectional survey and the National Longitudinal Survey (NELS:88) longitudinally followed students through secondary school from 1988 to 1992. NAEP national sciences surveys have been most recently carried out in 1986, 1990 and 1996. In 1986 they surveyed grades 3, 7 and 11 then switched in 1990 to grades 4, 8 and 12. NAEP analysis of hands-on science's relationship to test scores has been confined to grouping students by the level of hands-on science they report, averaging the test scores for each level and comparing them. In the 1986 NAEP, students in grades 3 and 7 were asked how often they used different types of equipment. Their responses were grouped into three categories: low, medium and high. Students in grades 7 and 11 were asked how often they were involved in innovative classroom activities including hands-on activities. Their responses were grouped into the same three categories. In both cases, the high response group scored significantly more than the low response group (Mullis and Jenkins 1988).

In 1990, students in all three grades were asked which of six broad categories of equipment or materials they had used. Their answers were grouped into four categories: none, 1 or 2, 3 or 4, and 5 or 6. For grades 8 and 12, the mean student scores monotonically rose with the response category with a significant difference between the none category and the 5 or 6 response (Jones, Mulllis, Raizen, Weiss, and Weston 1992).

In 1996, students in all three grades were asked if they had done hands-on activities or projects with seven different types of materials or instruments (eight for grade 12). Except for several cases involving 4th grade, students who answered yes had a higher group score than those answering no. Students who had used none of the materials or instruments scored lower than those who had used some in all grades. In addition, for grades 4 and 8 teachers reported how often they used hands-on activities with four possible responses: never or hardly ever, once or twice a month, once or twice a week, and almost every day. There was no difference in the student test score means for each of these groups in the 4th grade but in 8th grade the two highest response categories scored more than the lowest response category. The study further grouped students into science proficiency categories based on their test scores. The 8th grade students who reported more hands-on activities were more likely to be ranked at or above the Proficient level. The same was found for 12th grade students who self-reported higher categories of hands-on activities (Sullivan and Weiss 1999).

The NELS:88 included 25,000 8th graders in 1988 and about 11,000 were surveyed on science. Follow-up surveys were done in 10th grade in 1990 and 12th grade in 1992 with additional students added to maintain a nationally representative cross-sectional sample for each year. Horn, Hafner and Owings (1992) grouped test scores by answers to a teacher item concerning how often experiments were done in the classroom. Possible responses included never or less than once a month, once a month, once a week and almost everyday. The mean student scores for the categories of almost every day and once a week were significantly higher than those of the other two categories. Teacher reports on hands-on science done in the 8th grade were also found to have a positive relationship with a subset of the NELS 10th grade test dealing with quantitative operations, chemistry and scientific reasoning but not with subsets of the other test items (Hamilton, et. al. 1995). This positive relationship was not found in a similar analysis of 12th grade sub-scores (Nussbaum, Hamilton and Snow 1997) nor in analysis of the frequency of experiments and growth in test scores between 10th and 12th grades (Hoffer and Moore 1996). Another analysis of NELS subscores divided the test into life science and physical science scores. No relationship was found between either subscore and 8th grade student reports of taking a science course with a lab once a week but a positive relationship was found between the physical science subscore and an interaction effect for girls who reported yes (Lee and Burkham 1996). A further analysis of 10th grade students using both a student and a teacher reported hands-on composite variable found a positive relationship between the teacher reports and both subscores (Burkham, Lee and Smerdon 1997).

Up to this point, we have discussed research on the relationship of hands-on science and scores from multiple choice tests. Research on the relationship of hands-on work and performance assessments are few and inconsistent. In some cases, the analysis of the relationship was not done. For example, the First International Science Study included a hands-on test for 14 year olds in England and Japan and secondary school completers in England but the analyses did not correlate scores from these tests with the amount of hands-on science (Comber and Keeves 1973; Tamir and Doran n.d.; Kojima 1974). The 1986 NAEP included a pilot study using hands-on tests but the purpose was to assess these tasks rather than determine their relationship to hands-on science or other factors (Blumberg, et al. 1986). The current NAEP use both short and long answer constructive responses, some based on hands-on tasks, along with multiple choice questions but the past analyses use only the total score (Sullivan and Weiss 1999).

The results from research on the IEA are not conclusive. The second IEA international science survey included three hands-on tests for both Grade 5 and Grade 9 students. Each test was composed of three hands-on tasks to be done by a student in one class period. Correlations were reported for the relationship of hands-on science and hands-on test scores. For grade 5 student reports, the mean correlation was 0 (-.03 for the U.S.) and for grade 5 teacher reports the mean correlation was again 0 (the U.S. did not collect teacher reports). For the grade 9 student reports, the mean correlation was .09 while for the teacher reports it was .30 but this figure was skewed upward by one country (Doran and Tamir 1992).

A small scale study found that 5th grade students using a strongly hands-on curriculum had a mean score on a hands-on test one-half standard deviation higher than those using a textbook based curriculum but this result may be confounded since the student with the hands-on curriculum also scored higher on a cognitive ability test (Baxter, Shavelson, Goldman and Pine 1992). Only in the case of lab skill tests has hands-on science been shown consistently associated with higher test scores (Yaeger, Engen and Snider 1969, for a review see Hofstein and Lunetta 1982). This research has not determined how lab skills contribute to overall science achievement or achievement in specific domains.

Overall then, we do not find a consensus on the relationship between hands-on science and achievement. The experiment based research has not supported the relationship between hands-on science and student achievement. The curricula focused research has found a relationship with achievement but it is not clear whether hands-on science or some other aspect of the curricula are responsible for this. The survey based research provides mixed results.

Issues Surrounding this Research

The research on hands-on science has grown more sophisticated over time. However, there are still several issues that have yet to be resolved. These issues may in part be responsible for the lack of agreement in past research. Here we discuss four of these issues: 1) variables related to both achievement and hands-on science, 2) a lack of large data sets containing individual student performance test scores along with important explanatory variables, 3) a differential relationship of hands-on science to student achievement by student ability, and 4) the multiple facets of hands-on science.

There are a large number of variables that may affect achievement and need to be considered when analyzing test scores. Some of these are related to the level of hands-on science as well as achievement and if not controlled for, a spurious relationship between hands-on science and test score can result. This problem is especially true for correlational studies, such as those done with the NAEP, which did not control for any variables but may also be true for multivariate analyses as well.

Among the variables that past research has identified as linked to achievement, several may also be linked to the level of hands-on science. These include SES, prior science course taking, achievement level of the class, and student ability or past achievement. Research using both the NAEP and NELS:88 finds that higher SES students achieve higher scores (Jones, Mullis, Raizen, Weiss and Weston 1992; Horn, Hafner and Owings 1992). Research has also shown that higher SES students report a greater amount of hands-on science (Horn, Hafner and Owings 1992; Hoffer and Moore 1996) and we have obtained the same finding using the NELS data for 8th and 10th graders for both teacher and student reports (see Table 5-3). Similarly, the number of science courses a student takes affects their proficiency in science regardless of their SES, race or gender (Hoffer, Rasinski and Moore 1995; Madigan 1997) and the more science courses a student takes, the more likely they are to have hands-on science. Additionally, the academic level of the class and the academic ability of the individual student are related to both high test scores and higher levels of hands-on science (Hoffer and Moore 1996) and we have obtained the same finding in the NELS data for grades 8 and 10 (see Table 5-3) using both student and teacher reports as well as in the RAND data when using student reports (see Table 4-4). Research not controlling for these variables could produce results different from that which has.

Standardized testing used in large surveys and many of the smaller experiments and evaluations of curricula has relied primarily upon multiple choice items. These tests have many positive attributes including ease of administration, relatively inexpensive scoring, ease of standardization, good psychometric properties and the ability to provide coverage of content in a short testing period. During the 1990s the focus on multiple choice testing came under a strong critique for a variety of reasons. Relevant to our research is the criticism that multiple choice tests cannot cover the wide range of skills that should be evaluated. Critics argue that multiple choice can only test narrow content areas and skills – especially short-term recall of facts and basic process skills - but cannot address the broader abilities of critical thinking, evaluation and problem solving (Miller and Legy 1993). Resnick and Resnick (1992) conclude that multiple choice standardized tests ask for quick bits of information, not linked to each other and not used for a whole task such as solving a problem or interpreting result: in short, they are not linked to thinking. Madeus et al. (1992) reviewed six standardized tests for math and science (with copyrights of 1985-1991) and found them to focus on short term recall questions concerning facts, definitions, and applications of formula, all of which could be learned by rote. The tests covered little procedural knowledge or problem solving and reasoning. A broader critique is that multiple choice tests are irrelevant to daily life and that testing should replicate the challenges people will face in the real world (Wiggens 1989).

These critiques are important to our research as they imply that focusing on multiple choice testing may overlook some of the benefits of hands-on science. If we use multiple choice tests to stand for achievement and if these tests cannot assess some of the benefits of hands-on science (such as teaching procedural knowledge and skills) then we may not capture the actual relationship between hands-on science and tests. Murnnane and Raizen (1988) make the specific charge that multiple choice testing cannot pick up what is taught through hands-on science.

A key conclusion drawn by the critics is that multiple choice testing needs to be supplemented (and in some cases supplanted) by alternative forms of assessments covering a wider range of knowledge and skills. These assessments are of many types and can be categorized in a number of different ways (OTA 1992; Herman, Aschbacher and Winters 1992). They may include such tasks as short or long written responses, oral presentations, the actual performance of a task, or the results of a performance (e.g. an exhibition or portfolio of past work). For the evaluation of hands-on science, alternative assessments may require the combination of the performance of hands-on activities with written or oral means. The performance section would allow the assessment of the hands-on skills while the written or oral section would address other procedural skills of science (e.g., developing hypotheses, recording data, and making conclusions based on results) as well as content knowledge associated with the activities.

The criticisms of multiple choice testing and the promotion of alternative testing represent one train of thought rather than the accepted viewpoint. Both in theory and practice, multiple choice remains widely defended and used for standardized testing. Mehrens (1992) provides a literature review to counter many of the criticisms of multiple choice testing including its narrowing of content and skills tested and inability to test higher thinking skills. He notes, though, that there is general acceptance that some types of procedural knowledge cannot be tested using multiple choice but require a performance assessment.

In the past, only a small number of testing surveys included hands-on activities. Today, while there is more alternate assessment in use, much of it is not relevant for our work. In some cases, only writing items are used (e.g. commercial tests such as the Stanford Achievement Test 9). In other cases, hands-on activities are done but the reported data cannot be used (e.g., the 1999 NAEP mixed the constructive response scores with the multiple choice scores, the Maryland's MSPAP program does not provide individual student test scores, and data from the Third International Math & Science Study has not been released). There is then a paucity of data to assess the links between hands-on science and hands-on science assessment that has two implications for our research.

First, it is important to understand how hands-on science relates to multiple choice test results but it is also important to determine if the type of testing (be it multiple choice or some alternative form) changes this relationship. If so, we may be missing some connection between hands-on science and student achievement. Second, while performance tests may have greater face validity for testing hands-on instruction, they have not been proven to have more construct validity than multiple choice tests. Their justification on these grounds then remains open to question and analysis.

A third issue concerns how student ability affects the relationship of hands-on science with student achievement. Past research has examined certain interactions of hands-on science with specific student groupings, for example gender and race/ethnicity (Peng and Hill 1995; Lee and Burkham 1996; Burkham, Lee and Smerdon 1997). But there has been little work on whether the relationship differs for higher versus lower ability students. From a policy point of view, this interaction is important for the public acceptance of hands-on science. Methods of instruction are often pulled in two directions with regards to the issue of equity. On one hand, methods may be supported that have the goal of increasing equity by improving the performance of lower ability students. On the other hand, if these methods have adverse effects on higher ability students, their introduction may be opposed. Politically acceptable methods then must benefit both groups or benefit one group without harming the other (i.e., one group may achieve more under the different method of instruction but the other group must achieve no less under it than when using other methods). In the case of adverse effects on one group, the method may be restricted to programs strictly for the other group. As our analysis is concerned with the broad application of hands-on science across the majority of classrooms, we are most interested in determining whether we find a benefit in the use of hands-on science without an adverse effect for any ability group.

From a theoretical point of view, it is expected that hands-on science should have a differential effect for higher versus lower ability students. Current theory, though, provides support for a differential effect that could favor either high or low ability students. Research on the interaction of student aptitudes (including ability) and instructional methods (known as Aptitude-Treatment Interaction or ATI) has found that teaching methods often have differential effects based upon student ability (Cronbach and Snow 1981). Specifically, instructional methods that require greater responsibility of the learner benefit higher ability learners more while methods that pre-process the material better help lower ability learners. Hands-on science requires students to do more on their own and draw conclusions for this work. This requirement for greater student responsibility argues that hands-on science may be of greater benefit to higher ability students or lessor benefit to lower ability students. There is some evidence for this from analysis of the hands-on curricula developed in the 1960s (Koran & Koran 1984). If true, hands-on science would be politically less acceptable as it would have an adverse impact on lower ability students.

At the same time, ATI also proposes that higher ability students benefit from instructional methods based on abstract concepts (Koran & Koran 1984). Hands-on science's strength, conversely, is in making the abstract concrete through physical examples. From this point of view, hands-on science may better benefit the lower achieving student who is not yet ready for instructional methods based solely on abstractions and hinder the higher achieving student who is. Development theory would also support this view as hands-on science would help students move up through the concrete thinking stage of mental development to the higher abstract thinking stage. In this case, hands-on science may benefit lower achieving students but not higher ability ones

Time constraints may also contribute to a differential impact of hands-on science based on student ability. Hands-on science requires more time to cover the same material than other methods thereby reducing the time available to cover other material. Higher ability students may be ready to cover more material more quickly using other methods besides hands-on science, especially if they can grasp abstract concepts with a minimum of concrete examples. If this is true, then when taught using hands-on science they will cover less material in class and learn less. Because this is an issue of material coverage, any differential effect of hands-on science for higher ability students will more obviously appear in tests covering more material. If true, there could strong opposition to the general introduction of hands-on science by higher achieving students and their parents.

A fourth issue for researching hands-on science is its multiple facets of quantity, quality, and instructional approach. Quantity is the measure of how much hands-on science is done in the classroom and/or how often it is done. Quality describes the skill of the teacher providing instruction through hands-on science and the value of the activities used. Third, hands-on science can be taught through a number of instructional approaches, as detailed in Chapter 2. For both theoretical and practical reasons, research has focused on quantity measures.

The critical distinction between hands-on science and conventional science is students' hands-on activities in classroom. Students' involvement in hands-on activities exposes them to vivid connections from abstract concepts to concrete examples and from scientific principals to scientific applications, which contribute to students' better understanding of the scientific knowledge and their ability to solve problems using their scientific knowledge. Quantity of hands-on activities in classroom captures this distinction between hand-on science and conventional science instructional methods as well as the degree to which hands-on activities are used. Thus, operationalizing hands-on science using quantity is valid.

In practice, data collection has focused on quantity. Quantity issues are considered more reliably surveyed as the amount of hands-on science done in a classroom is considered memorable. Teachers' reports on hand-on amount are relatively more reliable since the amount of hands-on activities is included in their lesson plans. Students' reports are relatively more reliable since hand-on activities are more interesting and more likely to be remembered. This assumed reliability combined with cost issues has led to a focus on the use of surveys versus classroom observation or use of teacher and student logs. Teachers and or students are most often surveyed as to the overall amount or frequency of hands-on work activities they have done in class. In addition, they may be asked about the use of specific materials or instruments or the completion of activities on specific topics. Proxy questions assumed to be related to the level of hands-on science may also be asked, such as frequency of lab reports.

Surveys concerning behavior (such as the quantity of hands-on work) are subject to several sources of error, primarily memory, possible motivations to misreport, and failure to understand the question correctly (Sudman and Bradburn 1991). These sources of error can be reduced in the structure and administration of the survey. Memory problems can be reduced by making items very specific, covering longer time periods to avoid respondents compressing time (moving events forward into the time period reported on), and allowing respondents to review their records. Motivation problems can be reduced by using less threatening questions, allowing less specific answers (such as daily, weekly, or monthly), and using more anonymous methods in administration and recording. Problems of understanding can be reduced by using familiar words (Sudman and Bradburn 1991; Rossi, Wright and Anderson 1983). Surveys on the quantity of hands-on work use many of these techniques. Surveys have specific items on frequency of experiments, they cover fairly long specific time periods of one course, they allow teachers to review their records, they often use open answers, they are given with a promise of anonymity for both teachers and students, and they use words familiar to science classrooms (e.g. experiments, scientific equipment, and observations).

The impact of these sources of error can be identified using verification studies. The work that has been done specifically on verifying surveys on classroom instructional behaviors has supported the use of teacher surveys. Burnstein, et. al. (1995) found that surveys of mathematics teachers gave an accurate picture of their level of use of different instructional methods when compared against a five week teacher log but noted that this finding might occur because there seemed to be little variation in teachers' methods overall. Porter (1995) found a correlation of .65 between math and science teacher survey responses concerning the amount of lab work and their logs kept over a year. Verification of student reports of methods used in the classroom has not been done though students surveys have been found to have reasonable construct validity concerning grades, course taking and some family variables (Fetters, Stowe and Owings 1984; Valliga 1986).

Measuring the quality of the hands-on science has been constrained by two factors. First, there is little agreement on how to measure quality of instruction. Second, quality of instruction is more open to individual interpretation than a quantity value. A small scale study of algebra teaching that tried to validate survey responses using classroom observation found that a composite of survey items could separate teachers using certain practices from those who did not but could not differentiate the quality of the practices among teachers using them (Mayer 1999). To ensure reliability, classroom observation would be the desired method of measurement but the high costs of this approach have constrained its use. For these reasons, most studies of hands-on science have not been able to address the issue of instructional quality.

Measuring the instructional approaches to hands-on science also faces difficulties. First, the approaches outlined above are not always clearly defined, are expected to vary with student ability, and may be used in various combinations by teachers. These factors increase the complexity of data collection even for classroom observation. Some of the large-scale surveys, like the NAEP and NELS:88, asked questions regarding teacher's overall instructional approaches but have not specifically asked about approaches used with hands-on science. Small-scale experiments often compare hands-on science teaching versus other methods (e.g., lecture or demonstration - for example, Yager, Engen and Snider 1969). As a result, they compare the quantity of hands-on science (some set amount in one classroom versus none in the other classrooms) rather than the interaction of quantity and instructional approach.

From a policy standpoint quantity is the first factor to examine because it can be most easily increased. Increasing quality of instruction or changing the instructional approach would require a much more intensive professional development program for teachers due to the current mix in the quality of teaching and in the instructional approaches used. A strong relationship between quantity and achievement would support Bredderman's (1992) inference that quantity of hands-on work and/or a focus on process may have a greater effect on achievement than instructional approach. If no link is found, though, concerns will remain whether quality or instructional approach play an unmeasured role. Ongoing attempts to improve the ability to measure both quality and instructional approach in the classroom, through both observation, teacher logs, and surveys will lead to future research better able to account for links between quality and approach of hands-on science and achievement. These will be important to refining suggested uses of hands-on science, e.g. only with specific instructional approaches or when teachers can provide a certain level of quality activities.

Improving Upon Past Work

Our goal is to improve upon past research on the relationship of hands-on science and student achievement. We do this by addressing the issues discussed above. First, the RAND data set includes both multiple choice and performance test scores from the same students. The performance tests were based on actual hands-activities done by the students who in addition wrote up their results and answers to further questions requiring reflection on the results. In addition, the data set contains both teacher and student reports on the level of hands-on science in the classroom. Thus, the RAND data set contains the two types of assessments and the level of hands-on science needed for this research.

In addition, because the two types of tests were taken by the same students, we can analyze a further topic not much looked at in the literature. Assuming there is a positive relationship between hands-on science and test score, we can consider whether there is a stronger relationship between hands-on science and one of the two types of tests. Past research has noted the possibility of such a differential relationship (Doran & Tamir 1992) and proponents of alternative assessment and hands-on science assume it (Murnanne & Raizen 1988). We might predict such a result as process skills are often hands-on in nature and performance tests are thought to focus on process skills more than multiple choice tests but there has not been a direct test of it.

A further contribution of this comparison is to provide some evidence regarding the claims that performance tests are more valid indicators of what students learn through a hands-on approach. A differential relationship between hands-on science and the two types of tests would support this claim while the lack of one (or one in the wrong direction) would not.

Second, we can avoid many of the concerns for a possible spurious relationship between hands-on science and student achievement by using multivariate analysis and including a broader array of variables that may be linked to both hands-on science and achievement. This is especially true for our analysis of the National Educational Longitudinal Survey (NELS:88) which contains a wide variety of this type of variable. NELS:88 provides us with the SES of the students' families, measures of student ability or achievement, achievement level of the class, and prior science course taking. In addition, we create our own composite variables, discussed in Chapter 5, to better reflect our conception of hands-on science and carry out additional analyses to determine if the variables linked to hands-on science and achievement affect our findings. Much of this additional analysis focuses on course-taking because it has several dimensions such as type and order of course-taking.

Third, we will test for a differential relationship between hands-on science and student achievement by student ability. As noted above, we will be using measures of student ability or achievement in our analysis to avoid spurious results. We will look at the interactions of these ability measures with the levels of hands-on science to determine if ability affects hands-on science's relationship with achievement. We will first determine if these interactions should be included in our analysis. If we find they should be included, then we will determine if they actually produce a differential relationship.

Fourth, we will continue the focus on quantity as our measure of hands-on science for both theoretical and practical reasons. Theoretically, quantity measures the primary difference between hands-on science and conventional instructional methods that we wish to measure. Practically, quantity is easier to measure, has higher reliability and is more widely available. Using the quantity of hands-on science also allows us to estimate the appropriate amount of hands-on work. As discussed previously, because of time constraints too much hands-on activities will greatly reduce the coverage of scientific topics, leading to a negative relationship with achievement. Our analysis will include this possibility and we will attempt to determine if there is a level of hands-on science at which the relationship to test score diminishes.

It is important to note that while the amount of hands-on activities in the classroom captures the distinction between hands-on science and conventional science, it does not distinguish the variations within hands-on science. Different instructional approaches to hands-on science (e.g., discovery or inquiry) and the quality of instruction (e.g., quality of the kit, teachers' preparation, and teachers' ability to stimulate the classroom) affect students' achievement just as they do conventional instructional methods. Our measure of quantity of hands-on science will not capture these variations within hands-on science. It would be ideal if we could assess various instruction approaches and quality of hands-on science by comparing various instruction approaches and quality of conventional methods of instruction. However, we lack data describing instruction approaches and quality for either hands-on science or conventional methods. Thus our estimate of the relationship of the quantity of hands-on science with achievement will capture the average effects of various instruction approaches and the quality variation in hands-on science in comparison with the average effects of various instruction approaches and the quality variation in conventional instruction methods.

Hypotheses

Our analysis will center on the testing of three hypotheses regarding the relationship of hands-on science and student achievement. More specifically we will examine the relationship of the quantity of hands-on science with student scores on standardized tests.

Our first hypothesis concerns the expected positive relationship between hands-on science and achievement. We hypothesize that students engaged in more hands-on science will score better on standardized tests than students who carry out less hands-on science, all other things being equal. Ideally, we hypothesize that a student's test scores would improve if they received greater hands-on science. As we cannot practically test this hypothesis, we have operationalized it by predicting that students who have received more hands-on science will score higher on tests than their counterparts who received less . We derive this hypothesis from the different theories of learning relevant to hands-on science, all of which see a benefit in the provision of concrete examples of abstract knowledge. These theories propose an especially important role for real-life illustrations of content knowledge to help students pass through the concrete stage to the abstract of thinking in the development of their minds, or to improve information retrieval from long-term memory by creating greater associations between pieces of knowledge, and to help prevent or correct misorganization of this knowledge. For process knowledge, these theories propose several benefits of hands-on science. First, many process skills are hands-on in nature and can only be learned through hands-on practice. Second, some of these skills link abstract concepts to empirical reality (e.g. the concept of variables and how to identify them) and learners would benefit from concrete examples of them. Third, hands-on activities offer opportunities for applying all the process skills in combination while focusing on specific content knowledge.

Past research has found mixed results regarding this hypothesis. We hope to obtain more consistent results using two data sets. Using multivariate analysis with increased control of variables we will address additional factors that may be confounding the relationship of hands-on work and test scores. Furthermore, we will test the hypothesis using two different types of tests, multiple choice and performance. As each type of test measures achievement with particular emphasis on a different domain, content versus process knowledge respectively, a finding of a positive relationship between hands-on work and both types of tests would provide more robust evidence for the relationship.

From a policy perspective, a rejection of the null hypothesis (that there is no relationship between hands-on science and test scores) would provide evidence to favor the increased use of hands-on science, the continuation of the policy of the 1960s. Current attempts to temper the use of hands-on science and redirect it under an inquiry approach might better be held off until further research examined the interactions of hands-on science and instructional approach. Conversely, a failure to reject the hypothesis would support the tempering of its use unless supportive results were obtained regarding the roles of instructional approach or quality on its relationship to achievement.

The second hypothesis grows out of measuring student achievement using the two types of tests. Specifically, we hypothesize that performance tests will reflect a greater difference in scores among students receiving different levels of hands-on science. Hands-on science benefits students taking performance tests because it teaches process skills required and assessed in performance assessment. Conversely, multiple choice tests are theoretically criticized for an inability to assess process skills and even their proponents note they it cannot cover all procedural skills.

Policy-wise the results from this analysis will have implications for the adoption of hands-on science and for standardized testing. As the majority of standardized testing is done through multiple choice tests, a stronger relationship with performance tests will identify a deficit in assessing for process skills. As a result, revising multiple choice tests or supporting wider use of performance tests may be justified. These revisions could lead to further support for the adoption of hands-on science. For as tests are revised, student test scores would be expected to show a greater difference between results by level of hands-on science (though a lesser difference between the same level by test type). Failure to find evidence in favor of this hypothesis, assuming that evidence had already been found for Hypothesis 1, would support the conclusion that both multiple choice and performance tests capture the benefits of hands-on science. This finding would provide less justification for revising multiple choice tests or increasing the use of performance tests (unless other research found that neither was well designed to capture process skills). With either finding, because of the debate over how to teach process skills, further research could be done on whether interactions of quantity of hands-on with instructional approach provide different results.

Third, we hypothesize that there is a differential effect of hands-on science for students of higher versus lower ability. While past research has identified similar effects for other types of instructional methods, the theoretical generalizations induced from that work is unclear on which way the differential effect will tilt. That hands-on science requires student to take more responsibility for their work should benefit higher ability students and we should see a differential effect in favor of higher ability students. On the other hand, that hands-on science makes the abstract more concrete should benefit lower ability students and we should see a differential effect in favor of lower ability students. Furthermore, the time requirements of hands-on science may reduce the number of topics a student is exposed to in class. For higher ability students able to cover a greater number of topics, hands-on science may reduce the amount they learn. This should be reflected in lower test scores, especially tests that cover a broad range of topics. Considering these opposite effects together, we hypothesize that hands-on science has a smaller positive relationship with achievement for higher ability students than lower ability ones. In particular, we would expect that any differential effect that favors lower ability students would be more strongly seen in a comparison of multiple choice test scores as multiple choice tests reflect a wider range of topics than performance tests.

Past research has not directly examined this interaction. Ability-interaction research has not looked directly at hands-on science. Research on hands-on science has focused on racial/ethnic and gender groups rather than ability groups. Although in some cases this work has controlled for ability, it has not examined the interaction of ability and hands-on science.

From a policy perspective, evidence in favor of the hypothesis of the existence of a differential impact by student ability (especially if one group found zero or a negative relationship) would make a general acceptance of hands-on science more complicated. Such a result would lend support for the differential use of hands-on science for only that group that benefited. The rejection of the hypothesis or if a weak differential was found (e.g., positive association with achievement for all ability groups though greater for one) would support the widespread use of hands-on science.

Data Sources

To tests these hypothesis we will analyze two separate data sets based on surveys of students and teachers. Both data sets include information on the level of hands-on work, student test scores, and other student and school covariates. The level of hands-on science is provided by both teacher and student reports allowing us to compare results from both. In both surveys, students took standardized science tests and the scores are used to represent their level of achievement in science. As noted in Chapters 4 and 5, these tests exhibit the same characteristics of other tests used for this purpose. Students of higher ability and social economic status produce higher test scores while minority students (Black and Hispanic) have lower scores. The surveys also contain additional information on the students, their families and their schools that may be related to both hands-on science and test score. With them, we can use the hands-on science and test score data to perform a multivariate regression analysis and thereby obtain estimates of the relationship of hands-on science to test scores holding the covariates constant.

Our primary data set is the RAND 1994 study on 1400 8th graders in Southern California. The study surveyed students and teachers about the amount of hands-on science they did. Students took both a multiple choice and a hands-on standardized test. In addition, some student demographics and abilities were also surveyed.

The RAND data will be used to test all three hypotheses (see Table 3-1). We can determine whether there is a positive relationship of hands-on science to achievement

(Hypothesis 1) using both multiple choice and performance tests score. If so, we next can test whether the relationship is stronger for performance tests (Hypothesis 2). Third, we can test for a differential relationship depending upon ability level of the student (Hypothesis 3) using both multiple choice and performance test scores.

Table 3-1: Data Sets Used in Testing the Three Hypotheses

Data Set	Hypothesis 1	Hypothesis 2	Hypothesis 3
RAND	Using MC & PA Tests	Test	Using MC & PA Tests
NELS:88	Using MC Tests	Not Test	Using MC Tests

The advantages of the RAND data are the large student sample and test scores from both multiple choice and performance tests for the same students. However, the RAND survey has a small teacher sample and a limited number of covariates for the variables we are concerned may be linked to both hands-on science and achievement. Specifically, the RAND data contains an ability measure for students. The issue of past course-taking is less important as 8th grade students normally have taken a similar number and type of courses.

The National Educational Longitudinal Survey (NELS:88) will allow us to check and extend some of the results from the RAND analysis. NELS:88 followed a nationally representative sample of about 25,000 8th grade students in 1988 to 10th grade in 1990 to 12th grade in 1992. Students and teachers were surveyed on the level of hands-on science in the classroom. Students took a standardized multiple choice science test. Data were also collected on student demographics, past course work, ability, classroom level and school characteristics.

The NELS data will be used to test the first and third hypotheses. Multiple choice test scores will be used to determine whether there is a positive relationship between hands-on science and student achievement (Hypothesis 1) and whether there is a differential relationship for higher and lower ability students. These hypotheses will be tested for each of the three grades.

Like the RAND data, NELS:88 contains a large sample of students. Another advantage is its large sample of teachers. In addition, NELS:88 also contains more of the covariates that may be linked both to hands-on science and test scores. In this way, we can extend our analysis of the RAND data to see if the results are robust when these covariates are controlled. Further, we can extend our analysis to the upper grades and here. With this extension, two points must be considered. First, one of the theoretical rationales for hands-on science is the provision of concrete experiences in support of understanding abstract concepts. As students mature, they are better able to directly understand abstract ideas and require fewer concrete demonstrations (Piaget 1973; Gage and Berliner 1994; Lawson 1995). Under this theory, we would expect the relationship of hands-on science to test score to be weaker for older students. Comparisons of the 8th grade students from the RAND survey with the older NELS students, especially the 12th graders, may illustrate this theorized difference in the usefulness of hands-on science. Second, in middle school most students take the same science courses while in high school there is greater student variation in science course taking. For analyses using students in higher grades, more attention will need to be given to the covariate of past course-taking.

NELS:88 has some disadvantages as well. First, its 8th grade student survey lacks the items necessary to develop a student measure of hands-on work and its 10th grade survey lacks a key item. This constrains the value of comparing the two data sets when considering student reports of hands-on science. Second, it does not contain performance test scores among the three waves of data. This lack prevents us from testing Hypothesis 2 and focuses our testing of Hypotheses 1 & 3 on multiple choice tests. While we will not be able to draw any conclusions on performance tests and hands-on science and their differential relationship versus multiple choice tests, we will have further evidence on the links between hands-on science and the most widely used, standardized, measure of student achievement.

As noted earlier in this chapter, NELS:88 has been used to examine the relationship of hands-on science to multiple choice test scores. This study complements and extends that work. Through the use of multivariate analysis and a broad array of variables that may be linked to both hands-on science and achievement we extend those studies that lacked these characteristics while complementing, and possibly confirming, those studies which have them. Second, the bulk of the NELS research operationalized the concept of hands-on science using variables with strong face validity. This work develops a composite hands-on science variable using both face validity and factor analysis. By better addressing the measurement error in hands-on science variable we have greater confidence that we are indeed measuring the underlying construct of hands-on science and reduce the bias in the coefficient estimating the relationship of hands-on science and test score. Third, this study will include the teacher and student measures of hands-on science available from each of the three years of data whereas the past studies did not. Fourth, this study will examine how ability may affect the relationship of hands-on science and test score, an issue to which the NELS data has not yet been applied.

Thee following two chapters describe the analysis using each data set, RAND and NELS:88, respectively. Each chapter discusses the data set in greater detail including the measures of hands-on science and test scores, along with the covariates. Each chapter provides a descriptive analysis of the data, a description of the models used in the analysis, the results from the models, and tests of robustness of these results.

Chapter 4: Analysis of the RAND Data

Introduction

In the Spring of 1994, RAND collected data on 1384 8th grade students from 44 classrooms in 8 schools across 4 districts within the Los Angeles metropolitan area. Within each school, data were collected during students' science class over a one week period. Three types of information were collected: 1) the level of hands-on science done in the classroom from both teacher and student surveys, 2) student achievement based on student test scores including one standardized multiple choice test and two performance assessments, and 3) student characteristics from teacher reports.

To test our three hypotheses, the analysis of the RAND data will first examine whether student hands-on science in the classroom has a positive relationship with standardized science test scores (both multiple choice and performance assessment test scores). If such a relationship is found, further analysis will examine 1) whether this relationship differs by test type and 2) whether this relationship differs by student ability level.

Below we first discuss the measures developed from the three types of information collected and provide descriptive statistics regarding them for our sample. Next, we discuss the three models we will use to analyze the data. We report the results from these models and test them for their robustness. We end the chapter with a summary of our findings.

I. Measures

From the information gathered through survey and testing, we developed three types of measures. From the teacher and student surveys, we created scales to measure the quantity of hands-on science in the classroom. From the testing, we obtained measures of student achievement using both multiple choice and performance assessments. From teacher reports, we obtained student characteristics regarding race/ethnicity and student ability.

The Level of Hands-on Science

Teachers and students responded to different sets of questions regarding the level of hands-on science done in the classroom. Teachers were asked about the frequency of using materials and equipment and use of class time. Students were asked about performing specific types of hands-on activities and the number of times experiments were done under different formats. Teachers answered the questions on their own time while students were given a class period to complete them. From these responses separate teacher and student scales were created to represent the level of hands-on science done in each classroom.

1. The Teacher Scale

The RAND teacher survey includes 19 items of two types regarding hands-on science (Table 4-1). Ten items concern the frequency of use of different materials and equipment. Two of these, calculators and computers, are not necessarily relevant as they may or may not be used during hands-on science. Seven items cover materials and equipment often used in science class and the tenth item is a catch all for use of any type of materials or equipment. The strength of the first nine items is that they are very specific and so help the teacher focus on what was done in the classroom. Correspondingly, their weakness is their specificity as well. As not every type of material and equipment can be listed, classes using other materials and equipment will be undercounted by this approach. The tenth addresses the problem of classrooms using other types of materials while also giving one response where all teachers can identify their overall use of hands-on materials. At the same time it suffers from being less specific than the other items of its type.

The other nine items address the percent of class time spent on specific tasks. One of these directly addresses the frequency of hands-on science, "supervising labs in which students do experiments". Several of the other items may also be related. As student experiments may be done in small groups, require individual instruction and may include teacher demonstrations, these items may be positively related to the frequency of hands-on science. Conversely, the frequency of other instructional approaches (such as "providing instruction to the whole class") and non-instructional activities may compete for time with student hands-on activities and so may be negatively related to the frequency of hands-on science.

A single teacher scale was constructed for two purposes. Analysis using individual items would introduce a potential problem of multicollinearity and reduce the model's parsimony. A single scale avoids these problems and allows for data reduction. More importantly, a scale reduces the measurement error that can occur with individual items analysis. Factor analysis was used to identify the items providing the main source of variance in the scale[1].

Seven items loaded on the first factor[2]:

1) Frequency of calculator use
2) Frequency of use of weights, scales and balances
3) Frequency of use of flasks, test tubes and chemicals
4) Frequency of use of any equipment or materials
5) Percent of class time supervising student experiments
6) Percent of class time providing instruction to class as a whole - e.g. lecture (negative contribution to scale)
7) Percent of class time doing school activities not related to subject (negative contribution to scale)

The teacher scale was created by combining the response to each of the seven items and then calculating the average which was then applied to all students in the class. It can take on a value of 1 to 6 with one meaning the lowest level of hands-on science done in the classroom and six the highest[3]. The teacher scale has reasonable reliability with a Cronbach's alpha of .90.

One concern we have with the teacher scale is the small number of teachers, eighteen teachers teaching 44 classes, taking part in the study. With such a small sample,

[1] A principal factor solution was used with the first factor having a proportion of .3021. I used a promax method for rotation as there is no reason why factors would be orthogonal. 2-6 factors were tested to ensure consistent loading of the items in the scale on the same factor.

[2] Items 1 to 4 concern the types of materials actually used in class and therefore the greater the frequency of their use the greater the amount of hands-on science done (with item 4 being a catch-all regarding any use of materials). The lack of inclusion of similar items in this scale (such as frequency of use of batteries) may reflect their absence in the curriculum (e.g., electricity may not be covered in the 8th grade curriculum). Items 5-7 concern percent of class time on specific tasks. Item 5 is a direct measure of time spent on student hands-on science while Items 6 and 7 appear to compete with time spent on hands-on science as they make a negative contribution to the scale.

[3] The values of Items 6 and 7, which originally made a negative contribution to the scale, were reversed (by subtracting them from 6 – the highest possible score on the scale) so that they make a positive contribution and give a clearer view of differences in this scale among groups when viewed in the descriptive data.

we might expect low variation in teacher responses resulting in low variation in the teacher scale which would weaken our statistical power to identify the relationship of the scale to test score.

2. The Student Scale

The RAND student survey includes 9 items of two types which address hands-on science (Table 4-2). Four items address specific hands-on activities. Five items concern the number of times the student had done experiments in class under different formats.

A single student scale was constructed using the same method as for the teacher scale[4]. Three items concerning the number of times experiments were done load on the first factor[5]:

1. Did experiments where I was told all the steps to follow
2. Work with one or more lab partners to do experiments
3. Did experiments where I used scientific equipment, such as magnifying glass, graduated cylinder or balance

The student scale was made by combining the responses to each of the three items and then taking their average for each student. For any student, the student scale can take on a value of 1 to 5 with one meaning the lowest level of hands-on science done in the classroom and five the highest. The student scale has reasonable reliability with a Cronbach's alpha of .82.

[4] The first four items were converted into dummy variables (0 = never did this activity before and 1 = had done this activity before). A factor analysis was done on these 9 items (principal factor solution) and the first factor had a proportion of .8556. Rotation was done using the promax method as there was no theoretical reason why factors would be orthogonal and 2-6 factors were tested to ensure consistency of results.

[5] The lack of inclusion of the other two similar items in this scale may reflect that experiments in middle school are often done in groups (to teach cooperative work and to save on material costs) and have the steps laid out in detail.

In addition to the student scale, two student items were retained for the analysis because they ask specifically about topics covered in the two performance assessments. These items are:

1. Measured the lifting power of levers
2. Classified different things (such as plants, animals or materials) into groups

Two variables are constructed based on these two items and are called "lever dummy" and "classification dummy". They take the value of 1 indicating experience with the activity or 0 if otherwise. Because these dummies are based on single items, we have less confidence in the estimates of their relationship to test score. Due to measurement error, these estimates may be attenuated.

3. Imputation of Student Hands-on Science Scale

A substantial proportion of students did not answer all items on the survey creating missing data in regards to the student scale or dummy variables[6]. Students missing the scale were found to be missing all items making up that scale. Table 4-3 reports the number of students with a missing scale or dummy who have an available test score.

Missing data was addressed through imputation based on regressing the scale or dummy variable on the students' race/ethnicity, gender and ability rank. When describing the data, pre-imputed data is used. For the hypothesis testing, both the pre-imputed data and the imputed data were used. As the results were similar, the findings reported are based on the imputed data.

[6] No data was missing for the teacher scale.

Student Test Scores

Students took one multiple choice test and two performance assessments. The Iowa Test of Basic Skills (ITBS) multiple choice science test (Level 14, Form K) containing 46 items and 46 total points possible was taken within the publisher's recommended time. Exercise administrators trained by RAND administered the test. The classroom teacher remained in the room but was not involved in the test administration. The publisher computer scored the test. 1238 students have ITBS test scores.

Students took two performance assessments developed by RAND. The lever test focuses on whether the length of the lever or the relative position of its fulcrum have an effect on the force needed to lift an object. The lever test contains 7 items with 14 points possible. 1242 students have lever test scores. The other performance assessment requires students to develop a two-way classification scheme for a set of objects and then fit an additional object into that scheme. This classification test contains 13 items with 48 points possible. 1231 students have classification test scores.

The exercise administrators gave both performance assessments to each class during a single science class period with the classroom teacher in the room. For each test, students were given the hands-on materials to be used and a test booklet containing directions, spaces where results were to be filled in, questions to be answered and the space for these responses. Cardboard partitions were used to reduce student interactions. Students carried out the instructed activities, recorded their results and answered the questions in the booklets. The booklets for each test were scored by a team of readers, primarily science teachers, who were trained and supervised by RAND staff in using semi-analytic scoring rubrics. Booklets were separated into batches and each batch held only one booklet from a classroom. Batches were randomly assigned to readers and readers were blinded to student characteristics. Inter-reader reliability for scores was high (.95) for both tests. For more detail on the make-up of the performance tests, their administration, and scoring see Stecher and Klein (1995).

Student Characteristics

While students were taking the tests and survey, teachers were asked to list the students enrolled in their class and identify three characteristics of each: gender, race/ethnicity and ability rank. Teachers listed whether each student was a male or female, with 1377 students identified. Teachers categorized each student as Asian, Black, Hispanic, Other, or White. Race/ethnicity is identified for 1376 students. Only 37 students were categorized as Other which are too few to maintain as a separate category nor is there reason to include them in another category.

Teachers were asked to rank the general ability of each student relative to all students in the 8th grade at that school. They did so by assigning an Ability Code (1 to 5) for each student which placed each student in grade-level ability quintile. Ability Code 1 indicates the bottom 20% and 5 indicates the top 20%. For the rest of this work, this characteristic is called Ability Rank. Ability rank is identified for 1376 students.

Using the race/ethnicity data provided by the teachers, the classroom percent minority (the percent of non-White students in each class) was generated as a classroom level variable to be used as a contextual variable in the analysis.

II. Descriptive Statistics

This section reports descriptive characteristics of the sample. First, we provide the distribution of the students by gender, race/ethnicity and ability rank (Table 4-4). The means and standard deviations of the hands-on scales and tests scores are presented for the entire sample and by gender, race/ethnicity and ability rank. Statistically significant differences within each of these subgroups are noted. In Table 4-4, the superscripts attached to the means identify which subgroups within the same category are statistically significantly different. For example Panel 2 provides data by race/ethnicity and shows a mean teacher scale for Asian of 3.35: the superscript "B" above the mean shows that it significantly differs from the mean for Black which is 3.10 .

Second, we examine between and within class variation in the student hands-on scale. Total variation is composed of the variation in class means (between-class variation) and the variation in student deviations from the class mean (within class-variation). Descriptive statistics are given for both components of the variance and the relative contribution of each to the overall variance in the student scale is described (Tables 4-5 to 4-7).

Subgroups of the Sample

The three subgroups identified in this sample are sex, race/ethnicity and ability rank. Of the 1384 students, 52% are female and 47% are male (Panel I, Table 4-4). White students make up 38% of the sample, Hispanic are 26% followed by Asian with 18% and Black at 15% (Panel II). Concerning ability rank, teachers placed 18% of their students in the lowest quintile, 16% in the second quintile, 22% in the third quintile, 24%

in the fourth quintile and 20% in the 5th (or top) quintile (Panel III). This shows a bit of overestimating ability rank by teachers as the bottom two quintiles contain less than 20% apiece of the students.

The Teacher and Student Hands-on Scales

The descriptive statistics show greater variation in the student scale than the teacher scale. The teacher scale contains more values than the student scale (1-6 versus 1-5) yet it has a smaller standard deviation (.84 versus 1.18) (Panel I, Table 4-4). In addition, significant subgroup differences in the teacher scale are restricted to the race/ethnic subgroups while for the student scale they include both the race/ethnic subgroups and the ability rank subgroups (Panels II & III).

This greater variation is expected because of the difference in sample sizes of teachers versus students. With 44 classrooms taught by 18 teachers, the teacher scale is expected to have low variation. While the 1400 students are in the same classrooms with these teachers, the variation in their reports can be greater for a number of reasons including their actual participation in the activities, their perception of what an experiment is, and their memory.

1. The Teacher Scale

Teachers report an average hands-on science scale of 3.29 (on a scale of 1-6) (Panel I, Table 4-4). We find no difference between the average report for teachers who teach females and the teachers who teach males. This is true as well for teacher reports by ability ranks. In part, these results may be due to mixed gender and mixed ability rank in the classroom or by a failure to report differences in tracked classes taught by the same teacher even though separate surveys were filled out for each class. However, we find differences by race/ethnic group (Panel II). Compared to Black students, Asian and White students tend to come from classrooms where the teachers report greater hands-on science.

2. The Student Scale

Students report an average hands-on science scale of 3.48 (on a scale of 1-5) (Panel I, Table 4-4). There is no difference between reports for female and male students but differences do exist among ethnic groups and ability ranks. White students report significantly higher levels of hands-on science than Black and Hispanic students and Asian students report more than Hispanic students. Students classified in the two highest ability rank quintiles report more hands-on science than those in the other three quintiles.

For the two student items (Lever and Classification), 64% of students reported using levers and 86% reported classifying objects (Panel I). The only significant difference in reports among subgroups is that students of the highest ability rank report greater use of classifying than students in the lowest ability rank.

Test Scores

High variation in test scores appeared among subgroups, on the whole, in an expected manner justifying the need to control for these subgroups in the analysis. Whites and Asians scored significantly higher than Blacks and Hispanics (Panel II, Table 4-4). Students of high ability rank scored significantly higher than those of lower rank (Panel III, Table 4-4). Only in the case of the Classification Performance test do we see a difference by sex with female scoring higher than male.

1. Multiple Choice Test

The average ITBS score was 23 points out of a possible 46 points. We found significant difference in ITBS score within race/ethnicity and ability rank subgroups but not by gender. White students scored higher than Asian students (by 3 points) and Hispanic and Black students (by over 6 points). Asian students scored higher than Black and Hispanic students by over 3 points. Students in the highest and second highest ability ranks scored higher than students in all lower ability ranks. Students in the highest ability rank scored 7.5 points more than those in the lowest ability rank. Students in the middle ability rank scored higher than those in the lowest ability rank by 2 points.

2. Lever Performance Test

The average Lever score was 6.8 out of a possible 14 points. Lever scores were similar for males and females but differed significantly by race/ethnicity and ability rank subgroups. White and Asian students scored higher than Black and Hispanic students by about 2 points. Students in the highest and second highest ability ranks scored higher than students in all lower ability ranks. Students in the highest ability rank scored 3.5 points more than those in the lowest ability rank.

3. Classification Performance Test

The average Classification score was 28.4 out of a possible 48 points. Females scored higher than males by almost 2 points. White students scored higher than Asian students by 3 points and higher than Black and Hispanic students by about 10 points. Asian students scored higher than Black and Hispanic students by over 6 points. Students in the highest and second highest ability ranks scored higher than students in all lower ability ranks. Students in the highest ability rank scored 11 points more than those in the lowest ability rank. Students in the middle ability rank scored higher than those in the lowest ability rank by almost 4.5 points.

Within and Between-Class Variation in the Student Scale

Because every student in each class reported on the level of hands-on science, the total variation in the student scale can be broken down into between-class and within-class variation. Total variation measures the variation in student reports. Between-class variation measures the variation of the classroom means. Within-class variation measures the variation of the student within each classroom.

The purpose of breaking total variation down into between and within-class variation is to create a more valid student scale. Within-class variation represents student reaction to the level of hands-on science done in the classroom which may be caused both by actual differences in student hands-on science (e.g. through absences and failure to participate) or by differences in student perception. Between-class variation, then, may better represent the real instructional differences in the amount of hands-on science occurring between classes. Between-class variation may also better correlate with the real amount of hands-on science in the class because the use of class averages may take away anomalies in individual student reports. We will do a separate analysis of the two sources of variation and their relationship to test score to test the sensitivity of our results.

Below we provide descriptive statistics on the between and within-class variation including the breakdown of these two types of variation for the student scale and two student items, the mean of classroom means, and the mean of the student variation from the classroom means by subgroup.

1. Breakdown of Total Variation

Table 4-5 reports the total variation and the between and within-class variation for the student scale and two student items. Only the student scale contains a large percentage of both between-class and within-class variation. For the two dummy variables, the majority of variation is due to within-class variation.

2. Between-Class Variation

As classroom mean is an aggregate statistic of student reports, we would expect its mean to be similar to the mean of student reports (Panel 1, Table 4-4) but with a smaller standard deviation. Table 4-6 shows this to be the case.

3. Within-Class Variation

The mean of the student deviation from the classroom mean totals zero by definition so there is no need to report it in a table. However we can report the means of student deviation from the classroom mean for specific subgroups. While these will tend toward zero, they can show significant differences among subgroups. Table 4-7 reports the mean of the student deviation from the classroom means of the student scale and the two items by subgroups. For the student scale there is no significant difference between male and female but there are differences among race/ethnicity and ability rank subgroups. White student deviations are positive and greater than Asian or Hispanic students showing that White students reported more hands-on science than their Asian or Hispanic classmates. The same is true for students in the top two ability ranks versus those in the second lowest ability rank. For the lever and classification dummies, there are no differences among the subgroups.

III. Models

We employ three models to examine the relationship of hands-on science with standardized science test scores. Model 1 tests Hypothesis 1 as to whether there is a positive relationship between hands-on science and test scores. Model 2 tests Hypothesis 2 as to whether hands-on science has a stronger relationship with performance test scores versus with multiple choice test scores. Model 3 tests whether our results regarding Hypothesis 1 hold when we break down our hands-on science measure into: 1) between-class variation and 2) within-class variation.

Each of these models is extended to include the interactions between hands-on science and ability ranks in order to test Hypothesis 3 that hands-on science has a weaker positive relationship with achievement for higher ability students.

The RAND data were collected using cluster sampling of students within classes. This clustering could cause the underestimation of the standard errors of the coefficients for the independent variables. The underestimation would be magnified for the hands-on science variable due to its classroom nature which would lead to overestimating the significance level of its association with test scores. We will use the Huber correction for our OLS models to produce robust standard errors so as to test our hypotheses more precisely (Huber 1967). The following discussion describes the three models (Models 1, 2 & 3) and their extensions (Models 1A, 2A & 3A).

Model I

To examine whether students who carry out more hands-on science score higher on standardized science tests controlling for student and classroom characteristics, we use the following regression model:

1) $Y_{ij} = \alpha_0 + \alpha_1 H_{ij} + \alpha_2 CL_j + \alpha_3 ST_{ij} + \varepsilon_{ij}$

where

Y_{ij} = the test score (multiple choice or performance test) for student i in class j.

H_{ij} = the level of hands-on science for student i in class j (for student reported data this includes the student scale and the lever and classification dummies: for teacher reported data this is the teacher scale).

CL_j = the class level variable (classroom percent minority) that may be related to test scores for class j.

ST_{ij} = the student characteristics (ability rank, gender, race/ethnicity) that may be related to test scores for student i in class j.

α's = the parameters to be estimated (α_1 being the vector of parameters we are most interested in).

ε_{ij} = the disturbance term for student i in class j.

Our regression is superior to the zero-order correlation method, which has been widely used in the literature. First, the model allows us to separate out the effects of other variables known to affect test scores (such as gender and race/ethnicity). The resulting regression coefficients will give us a more accurate measure of the relationship of hands-on science to test scores. Second, the regression estimates avoid the attenuation of the relationship between performance test scores and hands-on science that may occur in the correlation method. Because of the potentially lower reliability of our performance tests, correlations between hands-on science and performance test scores could be biased toward zero. We avoid this problem since regression coefficients estimated using the raw test scores do not depend upon the standard deviations of the test scores. Third, our model corrects for the effects of cluster sampling using the Huber correction.

The model will be estimated separately six times for all the combinations of the three different tests and the two different measures of hands-on science. The order of these six estimations is shown in Table 4-8.

Model 1 will be extended (to Model 1A) by adding terms interacting the hands-on science scales (teacher and student) with the ability ranks (symbolized by H^*AR_{ij}). Model 1A contains the four interaction terms of hands-on science and ability rank and takes the form:

$$\text{1A)} \quad Y_{ij} = \alpha_0 + \alpha_1 H_{ij} + \alpha_2 CL_j + \alpha_3 ST_{ij} + \alpha_4 H^*AR_{ij} + \varepsilon_{ij}$$

where, in addition to the symbols interpreted for Model 1:
H^*AR_{ij} = the 4 terms interacting the hands-on science scale and the top four ability ranks (the fifth term indicating the lowest ability rank is used as a reference) that may be related to test scores for student i in class j. The AR_{ij} variables themselves remain included under ST_{ij}.
α_4 = a vector of parameters to be estimated concerning the differences in the relationship of hands-on science and test scores by ability rank.

The parameters estimated for the interaction terms (α_4) will be added to the parameters estimated for the student or teacher scales (α_1) to determine the relationship of hands-on science to test score specifically by ability rank. For example, to calculate the relationship of hands-on science to ITBS test score for the top ability rank students, we will add the coefficient on the hands-on science scale to the coefficient of the interaction term between the scale and the top ability rank if both coefficients are significant. For the low ability group (which is the reference group), α_1 alone captures the relationship. These results will show if hands-on science's relationship to test score differs by ability rank.

Like Model 1, Model 1A is estimated separately six times for different test scores and teacher or student reports of hands-on science. Model 1A contains all the advantages of Model 1 while additionally offering a test of the potential differential relationship between hands-on science and test scores by ability ranks. To determine whether Model 1 or Model 1A is the more proper specification, an F-test will be performed to test the significance of the interaction terms.

Model 2

If we find that hands-on science is associated with test scores, we will test whether the relationship of hands-on science with performance test scores is significantly greater than the relationship of hands-on science with multiple choice test scores. One of the strengths of the RAND data is that the same students took both types of tests allowing us to compare differences in the coefficients for hands-on science estimated in the equations using performance versus multiple choice test scores. We do this by taking the difference between the two equations and we illustrate this approach with the four equations below. Eq. 1.1 is Model 1 for the performance test scores (all original subscripts have been removed for ease of reading and a subscript of "p" standing for performance test is used). Eq. 1.2 is Model 1 for the multiple choice test scores (a subscript of "mc" standing for multiple choice test is used.) Eq. 1.3 is the subtraction of Eq. 1.2 from Eq. 1.1. Eq. 2 is the specification for the resulting Model 2 (the subscript "d" standing for difference is used).

1.1) $Y_p = \alpha_{p0} + \alpha_{p1}H + \alpha_{p2}CL + \alpha_{p3}ST + \varepsilon_p$

1.2) $Y_{mc} = \alpha_{mc0} + \alpha_{mc1}H + \alpha_{mc2}CL + \alpha_{mc3}ST + \varepsilon_{mc}$

1.3) $Y_{p} - Y_{mc} = (\alpha_{p0} - \alpha_{mc0}) + (\alpha_{p1} - \alpha_{mc1})H + (\alpha_{p2} - \alpha_{mc2})CL + (\alpha_{p3} - \alpha_{mc3})ST + (\varepsilon_{p} - \varepsilon_{mc})$

2) $Y_d = \alpha_{d0} + \alpha_{d1}H + \alpha_{d2}CL + \alpha_{d3}ST + \varepsilon_d$

where
$\alpha_{d0} = \alpha_{p0} - \alpha_{mc0}$
$\alpha_{d1} = \alpha_{p1} - \alpha_{mc1}$
$\alpha_{d2} = \alpha_{p2} - \alpha_{mc2}$
$\alpha_{d3} = \alpha_{p3} - \alpha_{mc3}$
$\varepsilon_{d} = \varepsilon_{p} - \varepsilon_{mc}$

Model 2 will be estimated four times based on the combinations possible from: 1) two differences between test scores (Lever - ITBS and Classification - ITBS), and 2) two sources of hands-on science reports (teacher and student).

In Model 2, by taking the difference of equations, we can examine the significance levels of the difference between coefficients. Our focus is on the significance level of the coefficient for hands-on science, α_{d1}, which measures the difference between the coefficient for hands-on science in the performance test equation and the coefficient for hands-on science in the multiple choice test equation.

In interpreting coefficients from Model 2, each coefficient represents the subtraction of the coefficient from Equation 1.2 (which uses the multiple choice test scores) from the coefficient from Equation 1.1 (which uses the performance test scores). If these two coefficients were positive, a significant positive coefficient for their difference in Model 2 would show that the hands-on variable has a stronger relationship with performance test scores than with multiple choice test scores. If these two coefficients were negative, the difference in them would be a negative plus a positive (a negative minus a negative). In this case, a significant positive Model 2 coefficient would mean that the variable has a stronger negative relationship with multiple choice test scores than with performance test scores. Table 4-9 details how to interpret the signs of significant coefficients from Model 2.

Model 2 requires that the performance and multiple choice test scores be standardized so that they can be compared. Scores are standardized with a mean of 50 and a standard deviation of 10. Only students having both types of test scores will be included in the analysis. Model 2 maintains the benefits of using multiple regression by separating out the effects of other variables known to affect test scores and addressing the impacts of cluster sampling by using a Huber correction. The potential problem of low reliability of performance tests is not corrected by Model 2 because the standardization of the performance test scores will reflect their potentially greater variance.

Model 2 will be extended (to Model 2A) by adding terms interacting the hands-on science scales with the ability ranks (symbolized by H^*AR_{ij}). The interaction terms will be used to determine if the differential relationship varies by ability rank. Model 2A takes the form:

$$\text{2A)} \quad Y_d = \alpha_{d0} + \alpha_{d1}H + \alpha_{d2}CL + \alpha_{d3}ST + \alpha_{d4}H^*AR_{ij} + \varepsilon_d$$

Model 3

If Model 1 or 1A shows a relationship between hands-on science and test score, we can further examine this relationship for the between-class and within-class variation in the student scale.

Model 3 separates out the relationship of between-class variation from the within-class variation in hands-on science and test score using the form:

$$\text{3)} \quad Y_{ij} = \alpha_0 + \alpha_1 \overline{H}_{.j} + \alpha_2 (H_{ij} - \overline{H}_{.j}) + \alpha_3 CL_j + \alpha_4 ST_{ij} + \varepsilon_{ij}$$

where

$\overline{H}_{.j}$ = the classroom mean of the hands-on scale for class j (capturing the between-class variation).

$(H_{ij} - \overline{H}_{.j})$ = the student's deviation from the classroom mean of the hands-on scale for student i in class j (capturing the within-class variation)

The model will be estimated three times, once for each test. The raw test scores and the Huber correction will be used. We will use the estimated parameters, α_1 and α_2, to determine the relationships of between-class variation and within-class variation of hands-on science to test scores. As we have greater confidence that the between-class variation in the scale is a more valid indicator of the actual level of hands-on science we will have greater confidence in our Model 1 results if we also find α_1 to be positive and significant.

Model 3 will be extended to Model 3A by adding two sets of interaction terms: 1) the classroom mean hands-on science scale multiplied by the ability rank, and 2) the deviation from the classroom mean hands-on science variables multiplied by ability rank. The model takes the form:

$$\text{3A)}\quad Y_{ij} = \alpha_0 + \alpha_1 \overline{H}._j + \alpha_2 (H_{ij} - \overline{H}._j) + \alpha_3 \overline{H}._j AR_{ij} + \alpha_4 (H_{ij} - \overline{H}._j) AR_{ij} + \alpha_5 CL + \alpha_6 ST_{ij} + \varepsilon_{ij}$$

where

$\overline{H}._j AR_{ij}$ = the terms interacting the classroom mean of the hands-on scale for class j with the ability rank of student i in class j.

$(H_{ij} - \overline{H}._j) AR_{ij}$ = the terms interacting the student's deviation from the classroom mean of the hands-on scale for student i in class j with the ability rank of student i in class j.

The interaction terms will be used to determine if the between and within-class variation vary by ability ranks. The parameters estimated for the interaction terms will be added to the parameter estimated for the student scale to determine the overall relationship of hands-on science to test score by ability rank.

IV. Results

The results from the estimation of the three models and their extensions are presented in Tables 4-10 to 22. For all three models an F-test was done between the model and its extension to see if the interaction terms composed of ability rank and student hands-on scale jointly make a significant contribution. The interaction terms composed of the student items (Lever and Classification) and ability rank were not found significant and so were dropped from the extensions of the models.

For all three models, Ability Rank 1 and Ability Rank 2 (the two lowest quintiles) were found not to be significantly different from one another. They were combined and used as the reference group for ability rank. The same was found for the terms interacting them with the hands-on scales, therefore, the two interaction terms were combined as well. The Huber correction was used to adjust for cluster sampling of students by class.

Model 1: The Relationship of Hands-on Science and Test Scores

Model 1 and Model 1A test Hypothesis 1 that there is a positive relationship between hands-on science and test scores. We examine the coefficients on the hands-on scales (teacher and student) for each test score to test this hypothesis. Where appropriate the coefficients on the interaction terms composed of the scales and ability rank are added to the coefficient on the scale to determine the relationship by ability rank. These models explain nearly 30% of the variation in test scores.

No significant results are found for the teacher scale. In the case of ITBS scores, the coefficients from both models are statistically insignificant. For Model 1A the interaction terms for the higher ability ranks and the teacher scale are all non-significant (except for the Ability Rank 5 interaction term which is marginally significantly negative). In the case of the Lever and Classification scores the coefficients are insignificant as well.

We, therefore, focus the discussion of Models 1 and 1A on the student scale in the following order: 1) ITBS multiple choice test scores, 2) Lever performance test scores, and 3) Classification performance test scores. Findings for the covariates were similar when using the teacher and the student scales so these will be discussed with the results for the student scale.

The lack of significant results when using the teacher scale preempts the need to analyze Model 2 and Model 3 with the teacher scale. Therefore, the discussion of Models 2 and 3 will focus solely on results when using the student scale.

1. Multiple Choice Test Scores - the ITBS

Table 4-10 reports the results of Models 1 and 1A with ITBS scores as the dependent variable. The F-test of the inclusion of the interaction terms rejects the hypothesis that the two models are not different at the 1% significance level (see Appendix 4-1). Therefore, Model 1A is the appropriate model to use with the multiple choice test data.

The student scale shows a positive relationship with test score but the two dummy items do not. The coefficient on the student scale is a positive 1.28 and significant at the .01 level. The coefficient on the lever dummy is negative and non-significant while the coefficient on the classification dummy is positive and non-significant.

The interaction terms also show a significant relationship to test score and must be considered in the relationship of the hands-on scale to test score. The coefficients are negative and significant for the terms of the higher ability students: -1.50 for the term including Ability Rank 5 and −1.14 for the term including Ability Rank 4. The term including Ability Rank 3 has a marginal significant negative coefficient of -.88.

To determine the relationship of the scale to ITBS scores by ability rank, we combine the coefficient on the scale with the coefficient on the respective interaction term for each ability rank. The results are shown in Table 4-11.

Using a test of equivalence of coefficients (H_o: Coefficient on Scale + Coefficient on Interaction Term = 0), we find that we cannot reject the hypothesis that the final coefficients for Ability Ranks 3 - 5 equal 0 at the 5% significance level. In sum, we find a significant positive relationship between the scale and ITBS test scores for students classified in Ability Ranks 1 and 2 (the lowest two ranks). The failure to find a relationship for Ability Ranks 3-5 may be due to the lack of one or because we lack the power to find one. For this reason, we interpret the test results to mean that we find a relationship of near zero (rather than equivalent to 0) for students classified in Ability Ranks 3-5.

Concerning the other explanatory variables, we find significant negative coefficients for Female (-.89), Asian (-2.46), Black (-5.00), Hispanic (-2.66) and classroom percent minority (-6.46). We find that ability rank has significant positive coefficients that rise monotonically from Ability Ranks 3 to 5 (3.75, 7.52 and 10.64 respectively).

2. Performance Assessment - The Lever Test

Table 4-12 reports the results of Models 1 and 1A with Lever scores as the dependent variable. The F-test of the inclusion of the interaction terms using student reports does not reject the hypothesis that the two models do not differ with a significance level of 5% (see Appendix 4-1). Therefore, Model 1 is the appropriate model.

The student scale and classification dummy show a positive significant relationship to test score. The coefficient on the student scale is a positive .29 and significant at the .01 level. Surprisingly, the lever dummy is non-significant though the coefficient on the classification dummy is a marginally significant .60.

For the other explanatory variables, we find significant negative coefficients for Black (-1.72), Hispanic (-.97) and classroom percentage minority (-3.07). We find significant positive coefficients for Ability Rank 4 (1.58) and Ability Rank 5 (2.37).

3. Performance Assessment - The Classification Test

Table 4-13 reports the results of Models 1 and 1A with Classification scores as the dependent variable. The F-test of the inclusion of the interaction terms does not reject the hypothesis that the two models are not different (see Appendix 4-1). Therefore Model 1 is the appropriate model.

The student scale and classification dummy show a positive relationship to test score. The coefficient on the student scale is a positive .98 and significant at the .01 level. As expected, the coefficient on the classification dummy is positive and significant (2.57). The coefficient on the lever dummy is non-significant.

In addition, we find a significant coefficient for Female (1.23) and significant negative coefficients for Black (-6.72), Hispanic (-.5.01) and classroom percentage minority (-11.17). We find significant positive and monotonically increasing coefficients for Ability Ranks 3-5 (2.00, 4.30 and 6.89 respectively).

4. Summarizing the Results of Model 1 and Model 1A

Student reports of hands-on science in the classroom show a significant positive relationship with all three test scores. In the case of the multiple choice test, this positive relationship exists only for the students of lower ability rank while there is no relationship for students of higher and middle ability rank. In the case of both performance tests, this relationship exists for all students and there is no difference by ability rank.

The relationship of student and classroom characteristics, except sex, to test score are similar for all tests. Non-Asian minority status and classroom percent minority are associated with lower test scores. Higher ability rank is associated with higher test scores. Female is associated with lower multiple choice test scores and has a positive or no relationship with performance test scores depending on the test.

In order to compare the relative strength of the hands-on scale's relationship with the three different tests and to compare its relationship versus the student and classroom characteristics we can standardize the coefficients by dividing them by each test scores' standard deviation. This step introduces the potential problem of the attenuation of the estimated relationship with the performance tests as we use their standard deviations when standardizing the coefficients.

Table 4-14 reports the relationship of hands-on science to test scores as a proportion of each test's standard deviation. The coefficient for the student scale converts to 17% of a standard deviation for ITBS and 8% of a standard deviation for Lever and for Classification scores. This result gives the appearance that hands-on science is more strongly related to multiple choice test scores than to performance test scores. The coefficient for the classification item converts to 21% of a standard deviation for Classification scores greatly improving the overall relationship of hands-on science to performance test score for this test.

For Model 1A, appropriate when using the ITBS scores, the interaction of the scale with ability rank was found to make a significant contribution. The coefficients for the interaction terms for Ability Ranks 4 and 5 convert to -.15 and -.20 of a standard deviation for ITBS scores.

Concerning the other independent variables, the results for race/ethnicity are negative and fairly consistent in value ranging from one-fourth to two-thirds of a standard deviation (although Asian is only significant when using ITBS scores) as are those for classroom percent minority (which almost reaches one standard deviation). For Female, the coefficients convert to about one-tenth of a standard deviation when using the ITBS and Classification test scores. For ability rank, we see a monotonic rise that ranges from 49% to 139% of a standard deviation when using ITBS scores, from 44% to 65% for Lever and from 16% to 56% for Classification scores.

The coefficient in standard deviation units for the hands-on scale converts to a smaller figure than those of the covariates. The covariates, though, are on the whole dummy variables while the scale runs from 1 to 5. A shift from the lowest value to the highest value of the scale would convert to a four times larger percentage of a standard deviation which is about equivalent to what we see for the relationship of Hispanic or Black. Therefore, a higher level of hands-on science can offset the test score disadvantages of being Black or Hispanic, particularly for lower-ability students.

B. Model 2: Differences in the Relationship of Hands-on Science to Multiple Choice Versus Performance Tests

Models 2 and 2A test Hypothesis 2 that the relationship of hands-on science with performance test scores is significantly greater than the relationship of hands-on science with multiple choice test scores. The interpretation of the coefficients from these Models differs from the interpretation of coefficients from a traditional OLS model. Because these coefficients reflect the difference in a coefficient from Model 1 (1A) using performance test scores minus a coefficient from Model 1 (1A) using multiple choice test scores, their interpretation relies on the value and sign of the Model 1 (1A) coefficients.

Models 2 and 2A will also be used to test a finding we made using Models 1 and 1A. The interaction of Ability Rank and hands-on science was found to be significant for ITBS scores but not for the performance test scores. One difference, then, between the relationship of hands-on science and type of test is that the relationship differs by student ability for the multiple choice test but not for the performance tests. The coefficient on the interaction terms in Model 2A will provide additional evidence regarding this finding.

The results of the models are discussed below in the following order: 1) Lever versus ITBS, and 2) Classification versus ITBS. Within these two sections we first note the results for Model 2 as an aid in discussing the results from Model 2A. Model 2 assumes that the coefficient on hands-on science is the same for students of different ability. Because a significant interaction was found for ITBS scores using Model 1A, Model 2A is the proper specification and we focus the discussion on the results from it. The R^2 are low because these models are being used to examine the significance of the difference.

1. Differential Relationship of Lever versus ITBS Test Scores

Table 4-15 reports the results from Models 2 and 2A with Lever minus ITBS as the dependent variable. The results for Model 2 show no significant difference in the relationship of the hands-on science variables. The coefficients on the student scale, and Lever and Classification dummies are not significant. These results provide little evidence that hands-on science has a stronger relationship with performance test scores than with multiple choice test scores under the assumption that such a relationship holds the same for students of different abilities.

Since the hand-on and achievement relationship does differ by ability ranks for ITBS, we need to test Model 2A. For Model 2A, we again find that Lever dummy and Classification dummy are not significant. The coefficient on the student scale is the estimate of the difference in the relationship of hand-on science to lever versus to ITBS for students of lower ability and we find it insignificant, indicating a lack of difference. The coefficients on the interaction terms (specifically for the higher ability students) are also not significant. However, their sign is positive as expected if the relationship between hands-on science is not as strong for higher-ability students for the ITBS as it is for the lever test. The coefficient on the interaction terms in Model 1A was negative and near zero for Lever scores and negative and larger than zero for ITBS scores. Subtracting the latter from the former (a larger negative from a smaller negative) should give a positive result which is what we see in Model 2A. In sum, our results do not give us sufficient evidence for a differential relationship for hands-on science with different types of test score. However, this may be due to a lack of statistical power rather than the lack of such a differential as our coefficients on the interaction terms though not significant have the correct sign.

Among the covariates, significant positive differences were found for Female, Asian and Black. Based on results from Models 1 and 1A, we would expect that Female and Asian would show a positive difference because both Female and Asian had a significant negative relationship with the ITBS test score and a non-significant relationship (low positive for Female and low negative for Asian) with the Lever test score. In the case of Black, Model 1 showed a significant negative relationship with both ITBS and Lever test scores. The significant positive coefficient in Model 2A shows that Black has a stronger negative relationship with ITBS test scores than with Lever test scores. A similar positive value for the Model 2A coefficient was also found on Hispanic but was not significant.

2. Differential Relationship of Classification versus ITBS Test Scores

Table 4-16 reports the results from Models 2 and 2A with Classification minus ITBS scores as the dependent variable. For Model 2 we see no significant difference in the relationship of the hands-on science variables as the coefficients on the student scale, and Lever and Classification dummies are not significant. This again confirms that there is no differential relationship of hand-on science with classification vs. ITBS test if the assumption that such a relationship is the same for students of different ability holds.

Given that Model 1A is the appropriate model for ITBS, we proceed to test Model 2A. For Model 2A, the coefficients on the Lever and Classification dummies are again not significant. The insignificant coefficient on the student scale of hand-on science again indicates that there is little difference in the hand-on and test relationship for students of lower ability. However, we find a significant positive difference between the coefficients for the two interaction terms which include Ability Rank 4 and Ability Rank 5 (the higher ranks). These coefficients are of the expected positive sign. Their significance and sign provides evidence that the relationship of hands-on science with achievement is more negative for ITBS test scores than for Classification test scores among students of higher ability.

For the covariates, the coefficients on Female and Asian are significant and positive while the coefficients on Ability Ranks 4 & 5 are significant and negative. The results for Female and Asian were expected based on our Model 1 and 1A results. Female and Asian had a significant negative relationship with the ITBS test score. Female had a positive significant relationship with Classification test score and Asian had a non-significant negative relationship. Ability Ranks 4 & 5 showed positive relationships with both Classification and ITBS test scores when using Model 1 and Model 1A respectively. That their difference is significantly negative in Model 2A shows that the positive relation of higher ability rank to test score is stronger for ITBS than for Classification.

3. Summarizing the Results of Model 2A

The results from Model 2A show that Ability Rank does affect the relationship of hands-on science to test scores for the ITBS but not for the Classification test (the finding from the Lever test was indeterminate). This result can be considered in two ways. First, for higher ability ranked students, hands-on science is more positively related to performance test scores because higher ability students show a more negative relationship between hands-on science and multiple choice test scores. Second, for lower ability ranked students, the relationship of hands-on science to test score is not affected by type of test and so the relationship is neither stronger nor weaker for performance test than for ITBS test among students of lower ability. In sum, we find some evidence that the relationship is stronger for performance tests than for multiple choice tests for higher ability students only.

Regarding the other explanatory variables, the results reiterate some of the findings from Models 1 and 1A but also provide additional information. Like Models 1 and 1A, Model 2A shows that Female and Asian have a more negative relationship with multiple choice test scores than with performance test scores (and Female has a more positive relationship with Classification performance test scores). Model 2A provides new evidence that Black may have a less negative relationship with performance test scores than with multiple choice test scores. This was not found for Hispanic nor classroom percent minority. Model 2A also provides new evidence that higher ability ranks have a stronger positive relationship to multiple choice test scores than to performance test scores.

C. Model 3: The Relationship of Between-Class and Within-Class Variation in Hands-on Work to Test Scores

Models 3 and 3A are a further test of the results found with Models 1 and 1A. Total variation in the hands-on scale is broken down into between-class variation and within-class variation. Between-class variation, which may be a more valid measure of classroom hands-on science, is represented by the classroom mean of the student hands-on scale. Within-class variation, which arises from both classroom hands-on science and student perception of it, is represented by the student deviation from the classroom mean of the scale. If the results concerning between-class variation from Model 3 mirror those for total variation in Model 1, we will have more confidence in our finding of a positive relationship between hands-on science and test score.

The results from Models 3 and 3A are discussed in the following order: 1) ITBS score, 2) Lever score, and 3) Classification score. Model 3A originally contained two sets of interaction terms: 1) the classroom mean of the hands-on scale interacted with each ability rank, and 2) student deviation from the classroom mean of the hands-on scale interacted with each ability rank. The latter were not found significant individually (using a t-test) nor jointly (using an F-test) in all cases. Therefore, only the interactions of between-class variation with ability rank are included in the final model.

The results for the covariates were found to be similar to those of Model 1A for the ITBS scores and Model 1 for the lever and classification scores. Therefore, they are not discussed in detail.

1. ITBS Test Scores

Table 4-17 reports the results from Models 3 and 3A for ITBS scores. The F-test of the inclusion of the interaction terms rejects the hypothesis that the two models are not different (see Appendix 4-1). Therefore we use Model 3A.

Our results are very similar to those found using Model 1A. They show a significant positive relationship between the two hands-on scale variables and ITBS test score. The coefficient for the classroom mean of the hands-on scale is a significant 1.65 and for the student deviation from the mean a significant 1.16.

The interaction terms also show a significant relationship to test score and must be considered in the relationship of the classroom mean scale to test score. The interaction terms show a significant negative relationship for the terms including Ability Rank 4 (-2.52) and Ability Rank 5 (-3.73) and a marginal significant negative relationship for the term including Ability Rank 3 (-1.57). To determine the relationship of the classroom mean of the hands-on scale to ITBS test score by ability rank, we combine the coefficient on the classroom mean of the student scale with the coefficient on the respective interaction term for each ability rank. The results are shown in Table 4-18.

Using a test of equivalence of coefficients (H_o: Coefficient on Scale + Coefficient on Interaction Term = 0), we find that we cannot reject the hypothesis that the final classroom mean coefficient for Ability Rank 3 equals 0 at the 5% significance level and that we reject this hypothesis for the resulting final classroom mean coefficients for Ability Ranks 4 & 5. In sum, we find a positive relationship between the classroom mean and ITBS test score for Ability Ranks 1 and 2, a near 0 relationship for Ability Rank 3 and a negative relationship for Ability Ranks 4 and 5.

2. Within and Between Differences for Lever Test Scores

Table 4-19 reports the results from Models 3 and 3A for Lever test scores. The F-test of the inclusion of the interaction terms rejects the hypothesis that the two models are not different (see Appendix 4-1). Therefore Model 3A is the appropriate model.

The results show a positive relationship between the two hands-on scale variables and the Lever test scores. We find a significant positive coefficient for both the classroom mean of the hands-on scale (.87) and the student deviation from the mean (.39). The interaction terms have significant negative coefficients for the terms that interact the classroom mean and Ability Rank 4 and Ability Rank 5. The coefficients for the classroom mean scale by ability rank are shown in Table 4-20.

Using a test of equivalence of coefficients, we find that we cannot reject the hypothesis that the resulting final classroom mean coefficients for Ability Ranks 4 & 5 equal 0 at the 5% significance level. For students classified in Ability Ranks 1, 2 and 3 our final coefficient for the classroom mean hands-on scale is positive while for students classified in Ability Ranks 4 & 5, the coefficient is near 0. The positive relationship found in Model 1 holds for Ability Ranks 1 to 3 but does not for Ability Ranks 4 and 5.

3. Within and Between Differences for Classification Test Scores

Table 4-21 reports the results from Models 3 and 3A for Classification test scores. The F-test of the inclusion of the interaction terms rejects the hypothesis that the two models are not different (see Appendix 4-1). Model 3A is then the appropriate model.

We find significant positive coefficients for the two hands-on scales. The coefficient on the classroom mean scale is 2.17 and the coefficient on the student deviation from the mean is 1.65. We see significant negative coefficients for the terms that interact the classroom mean and Ability Rank 4 and Ability Rank 5. Table 4-22 shows the between-class relationship of hands-on science to Classification test scores by ability rank.

Using a test of equivalence of coefficients, we cannot reject the hypothesis that the resulting final between-class coefficients for Ability Ranks 4 & 5 equal 0 at the 5% significance level. Students classified in Ability Ranks 1, 2 and 3 have a positive coefficient for the classroom mean hands-on scale while students classified in Ability Ranks 4 & 5 have a coefficient near 0. As in the case of the Lever test score, the positive relationship found in Model 1 only holds for Ability Ranks 1 to 3 and not the higher two ability ranks.

4. Summarizing the Results of Model 3

The results from Model 3A show a positive significant relationship of hands-on science to all three test scores for lower ability rank students when using between-class variation, a better measure of the level of hands-on science. Within-class variation in hands-on science was found to be positively related to test scores for all students. This sensitivity analysis confirms the relationship between hands-on science and test score that we found in Models 1 and 1A. It also provides additional evidence that this relationship exists for students in lower and middling ability ranks but not for higher ability rank students, a finding we did not see in the analysis using Models 1 and 1A for performance tests. Our first findings in Models 1 and 1A may have been contaminated by the within-class variation of hands-on science which has a constant positive relationship with test score across type of test and ability rank.

V. Robustness of Results

To further examine the robustness of our results regarding the positive relationship of hands-on science and test scores and the role of ability rank in this relationship, we extend our models to check if we obtain similar findings. We focus on extending Models 1, 1A and 3A for which we had significant results (See Appendix 4-2 for a detailed description). We estimate two types of extensions. The first includes estimating the models separately for specific categories of students. The second extension involves checking for non-linearities.

The first extension addresses five specific categories of students:

1. Gender: female and male
2. Race/ethnicity: Asian, Black, Hispanic and White
3. Ability rank: Ranks 1 through 5
4. Within-class student agreement on the hands-on scale:
   a. Students from classes with low versus high variation in the scale
   b. Students with low versus high deviation from their class mean of scale
5. Students from classes with low versus high agreement with the teacher hands-on scale

For the first extension, two approaches are used. In the first approach, interaction terms between each of these categories and the hands-on scale are created[7]. One set at a time, these interaction terms are introduced into the original model which is then re-estimated. In the second approach, each model is estimated separately for each specific group.

[7] For Model 3A, the interaction included only the between-class variation in the hands-on scale as interaction terms using within-class variation were not found significant.

Results from both approaches are very similar to those of the original models. With few exceptions[8]: 1) the coefficient on the hands-on scale remains significant and positive, 2) the coefficient on the interaction terms remain negative for the two highest ability ranks, and 3) in the case of the first approach, the coefficients on the new interaction terms are largely not significant or do not change the major findings.

The second extension addresses possible non-linearities by including a squared hands-on scale term. Where appropriate, interaction terms between the squared term and ability rank are also included. When these models are estimated in all but one case the results do not support the use of the polynomial specification. In the one case of Model 3A when using Classification scores, we find evidence to support such a specification. The squared scale term has a negative significant coefficient and one interaction term composed of the squared scale and ability rank has a positive significant coefficient (the scale retains a positive significant coefficient and the interaction of the scale and the two highest ability ranks retain negative significant coefficients).

To follow up on this finding regarding Classification scores, we estimated Model 3 including the squared scale term separately for each ability rank and found that this term remains significant for the lower Ability Ranks 1-3. This provides some evidence that the scale's positive relationship to Classification test scores diminishes and turns negative at some level of hands-on science for lower ability rank students (in this case at 3.4 on the 1-5 scale).

[8] Results from the first approach using interaction terms were very similar to those of the original models. The second approach had a greater number of exceptions as the reduced sample sizes led to more insignificant coefficients and less precise estimates.

Overall, the results from these two extensions of Models 1, 1A and 3A show the robustness of the results from the original models. In almost all cases, we continue to see a positive relationship between the hands-on scale and test scores for lower ability students but not for higher ones. In the case of one test, Classification, results from our polynomial specification suggest that for the Classification tests there is an upper limit of hands-on work after which the positive relationship starts to decline.

VI. Summary

Figure 1 gives a schematic view of the analysis of the RAND data and the major findings. The analysis was carried out to test three hypotheses:

1. The level of hands-on science is positively related to student achievement as measured by standardized test scores, both multiple choice and performance.
2. The relationship of hands-on science to test score is greater for performance tests as compared to multiple choice tests.
3. This relationship is weaker for higher ability students.

Our results support the first and third hypotheses and provide little evidence for the second hypothesis. First, we find a positive relationship between hands-on science and both types of test scores when using the student reported hands-on scale. For multiple choice tests, this relationship occurs only for students classified in lower ability ranks. For students in higher ability ranks this relationship is zero or negative depending on whether we consider the classroom average report (Model 1A) or the individual student report (Model 3A). For performance tests, all students have a positive relationship when we use the whole scale (Model 1). When we use the between-class variation in the scale, the positive relationship remains for students of lower and middle ability rank but goes to zero for the students in the highest ability ranks.

For the second hypothesis, we found some evidence that the relationship is more positive for performance tests than for multiple choice tests for higher ability students but this is due to the significant negative relationship between hands-on science and multiple choice test scores for these students. For lower ability students, hands-on science has a positive relationship with test score regardless of test type.

Furthermore, we broke down variation in the hands-on scale into between-class and within-class variation. Breaking down the total variation in this fashion helps us to understand the differing roles of between-class variation which appears to be a clear measure in the difference in levels of hands-on science and within-class variation which may combine both students' engagement in their class's level of hands-on science and their own perception of this engagement. We find that the relationship between hands-on science and test scores is based on both types of variation. Within-class variation has a positive relationship with test scores for all ability ranks supporting the argument that greater student engagement in and perception of hands-on science is related to higher test scores. Between-class variation was found to have a similar relationship with test score as total variation for low ability students and no or a negative relationship for higher ability students.

Further analysis was done to test these results by looking at specific categories of students and checking for non-linearities. This work concluded that the original findings are robust. Analysis focused on individual categories of students led to the same results as those from the entire sample. For the majority of the models non-linearities did not occur. No upper limit of hands-on science was found except in the case of the Classification test.

The above results are obtained using the student reported hands-on scale. No relationship between hands-on science and test scores is found when using teacher reported data. Why the scale used causes such a difference in results is open to a number of interpretations. Only a small number of teachers were involved in this study and there was little variation in the levels of hands-on work they reported. This could be due to a lack of variation in instructional techniques or possibly because teachers who use low levels of hands-on science over-reported in response to current educational policy mandates to increase student hands-on science. On the other hand, perhaps, student recollections are somehow biased, for example better students report more hands-on science.

Table 4-1: Teacher Survey Items Regarding Hands-on Science

Item	Responses
How frequently do students in this class use the materials and equipment below during class time: 1. Calculator 2. Computer 3. Magnifying glass, microscope 4. Telescope, planetary models 5. Weights, scales, balances 6. Batteries, wires, bulbs 7. Flasks, test tubes, chemicals 8. Rocks, minerals 9. Dissecting tools 10. Any equipment or materials	1 = Never 2 = 1 or 2 times a year 3 = 1 or 2 times a month 4 = 1 or 2 times a week 5 = Almost every day
About what percent of time is spent in a typical week doing each of the following with the class: 1. Providing instruction to class as a whole 2. Providing instruction to small groups of students 3. Providing instruction to individual students 4. Demonstrating lab procedures or experiments to students 5. Supervising labs in which students do experiments 6. Administering tests or quizzes 7. Supervising field trips 8. Performing administrative tasks (e.g. taking attendance) 9. Doing other school activities not related to the subject	1 = None 2 = <10% 3 = 10-24% 4 = 25-49% 5 = 50-74% 6 = 75%+

Table 4-2: Student Survey Items Regarding Hands-on Science

Item	Responses
Before this week, have you ever done the following science activities: 1. Used litmus paper to see if a solution was an acid or a base 2. Used any method to measure the pH number of a solution 3. Measured the lifting power of levers 4. Classified different things (such as plants, animals or materials) into groups	1 = No 2 = Yes, in 8th grade science 3 = Yes, in another class 4 = Yes, both in 8th grade science and in another class
Before this week, how many times have you done the following activities in your 8th grade science class: 1. Did experiments where I was told all the steps to follow 2. Did experiments where I had to figure out several steps without the teacher's help 3. Work with one or more lab partners to do experiments 4. Did experiments by myself 5. Did experiments where I used scientific equipment, such as magnifying glass, graduated cylinder or balance	1 = Never 2 = 1-2 times 3 = 3-4 times 4 = 5-6 times 5 = 7 or more times

Table 4-3: Missing Hands-on Science Observations

Hands-on Science Scale	Students with Multiple Choice Test Score	Students with Lever Test Score	Students with Classification Test Score
Student Scale	147	135	132
Lever Dummy	155	144	141
Classification Dummy	154	143	140
n	1238	1242	1231

Table 4-4: Descriptive Statistics of Rand Data

Panel I: Total & By Gender

Variable	Total	Female	Male
Hands-on Science			
Teacher Scale	3.29	3.33	3.25
(1-6)	(.84)	(.86)	(.81)
Student	3.48	3.54	3.42
Scale (1-5)	(1.18)	(1.18)	(1.19)
Lever Dummy	.64	.64	.65
	(.48)	(.48)	(.48)
Classification	.86	.88	.85
Dummy	(.34)	(.32)	(.36)
Multiple Choice	23.11	22.91	23.33
Test	(7.68)	(7.32)	(8.06)
(46 points)			
Performance Test			
Lever Test	6.78	6.91	6.63
(14 points)	(3.62)	(3.59)	(3.66)
Classification Test	28.38	29.24^{M}	27.37^{F}
(48 points)	(12.24)	(11.90)	(12.53)
N	1384	725	652
% of sample	100	52.4	47.1

table 4-4 continued

Table 4-4

Panel II: By Race/Ethnicity

Variable	Asian	Black	Hispanic	White
Hands-on Science				
Teacher Scale	3.35^{B}	3.10^{A,W}	3.29	3.35^{B}
(1-6)	(.79)	(.69)	(.77)	(.96)
Student Scale	3.52^{H}	3.28^{W}	3.22^{A,W}	3.71^{B,H}
(1-5)	(1.18)	(1.08)	(1.23)	(1.16)
Lever Dummy	.63	.66	.63	.65
	(.48)	(.47)	(.48)	(.48)
Classification	.86	.87	.81	.91
Dummy	(.35)	(.34)	(.39)	(.28)
Multiple Choice Test	23.07^{B,H,W}	19.35^{A,W}	20.03^{A,W}	26.46^{A,B,H}
(46 points)	(7.02)	(7.08)	(6.65)	(7.42)
Performance Test				
Lever Test	7.3^{B,H}	5.28^{A,W}	5.35^{A,W}	7.94^{B,H}
(14 points)	(3.43)	(3.64)	(3.54)	(3.30)
Classification Test	29.75^{B,H,W}	23.23^{A,W}	22.60^{A,W}	33.11^{A,B,H}
(48 points)	(11.69)	(11.03)	(11.47)	(11.09)
n	250	213	354	522
% of sample	18.1	15.4	25.6	37.7

table 4-4 continued

Table 4-4

Panel III: By Ability

Variable	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Hands-on Science					
Teacher	3.19	3.36	3.33	3.34	3.23
Scale (1-6)	(.81)	(.75)	(.80)	(.91)	(.86)
Student Scale	3.25^{3,4,5}	3.27^{3,4,5}	3.59^{1,2}	3.60^{1,2}	3.59^{1,2}
(1-5)	(1.30)	(1.14)	(1.06)	(1.20)	(1.19)
Lever	.68	.71	.65	.61	.60
Dummy	(.47)	(.46)	(.48)	(.49)	(.49)
Classification	.81^{5}	.88	.87	.85	.91^{1}
Dummy	(.40)	(.33)	(.34)	(.36)	(.29)
Multiple Choice Test	19.5^{3,4,5}	20.5^{4,5}	21.76^{1,4,5}	24.96^{1,2,3,5}	27.04^{1,2,3,4}
(46 points)	(8.3)	(7.02)	(7.00)	(6.58)	(7.47)
Performance Test					
Lever Test	5.03^{3,4,5}	5.7^{4,5}	6.29^{1,4,5}	7.56^{1,2,3,5}	8.58^{1,2,3,4}
(14 points)	(4.01)	(3.48)	(3.20)	(3.26)	(3.11)
Classification Test	22.79^{3,4,5}	24.87^{4,5}	27.20^{1,4,5}	30.53^{1,2,3,5}	33.96^{1,2,3,4}
(48 points)	(12.87)	(11.91)	(11.46)	(11.52)	(10.59)
N	244	226	298	336	272
% of sample	17.7	16.4	21.7	24.4	19.8

Note: The superscripts show which groups significantly differ (p=.05 or less) from the group in question. For example, the mean for female classification test scores reads 29.24^{M}. The superscript notes that the mean significantly differs from the mean male classification test score.

Key to Superscripts

Gender:	F = Female, M = Male
Race/Ethnicity:	A = Asian, B = Black, H = Hispanic, W = White
Rank:	1 = rank 1 (lowest rank) to 5 = rank 5 (highest rank)

Table 4.5: Total, Between-Class and Within-Class Variation of Student Scale and Items

Student Scale or Item	Total Variation	Between-Class Variation	Within-Class Variation	Within-Class as a % of Total Variation
Student Scale	1.39	.56	.83	59
Lever Dummy	.23	.02	.21	91
Classification Dummy	.12	.01	.11	92

Table 4-6: Mean of Classroom Means of Student Scale and Items

Student Hands-on Science Scale or Dummy Variable	Classroom Mean
Scale	3.50 (.74)
Lever Item	.64 (.14)
Classification Item	.86 (.11)

Table 4-7: Mean of Student Deviation From Classroom Mean of Student Hands-on Scale by Student Characteristics

Student Characteristic	Hands-on Scale	Student Lever Dummy	Student Classification Dummy
Gender			
Female	.04 (.90)	-.01 (.46)	-.01 (.31)
Male	-.04 (.92)	.02 (.45)	.01 (.34)
Race/ethnicity			
Asian	-.07^{W} (.95)	.007 (.47)	.007 (.34)
Black	-.06 (.94)	.03 (.45)	.02 (.32)
Hispanic	-.14^{W} (.98)	-.02 (.46)	-.02 (.38)
White	.14^{A,H} (.82)	-.002 (.46)	.02 (.27)
Ability Rank			
Rank 1 (low)	-.07 (1.06)	.03 (45)	-.04 (.36)
Rank 2	-.20^{4,5} (.98)	.05 (.43)	.02 (.32)
Rank 3	.04 (.84)	.008 (.46)	-.001 (.32)
Rank 4	.08^{2} (.85)	-.04 (.47)	.01 (.34)
Rank 5	.08^{2} (.86)	-.02 (.48)	.04 (.28)

Note: Means reported with standard deviations provided in the parentheses

The superscripts above the mean notes which groups significantly differ (p=.05 or less) from the group in question. For example, under the Hands-on Scale, in the category for Race/Ethnicity, Asian reads -.07^{W} which means that the mean for Asian significantly differs from the mean for White.

Key to Superscripts

Gender:	F = Female, M = Male
Race/Ethnicity:	A = Asian, B = Black, H = Hispanic, W = White
Rank:	1 = rank 1 (lowest rank) to 5 = rank 5 (highest rank)

Table 4-8: Order of Six Estimations of Model 1

TEST SCORES			
Hands-on Science Scale	ITBS	Lever	Classification
Student Scale	1	3	5
Teacher Scale	2	4	6

Table 4-9: Interpreting the Sign of Model 2 (PA - MC) Coefficients

Sign of coefficient in Eq. 1.1 (Model 1: PA)	Sign of coefficient in Eq. 1.2 (Model 1: MC)	Difference of 1.1 – 1.2 (Model 2)	Resulting Sign of Coefficient in Model 2	Interpretation (assuming coefficient is significant)
+	+	(+) - (+)	+	Variable has a stronger positive relationship with PA scores
+	+	(+) - (+)	-	Variable has a stronger positive relationship with MC scores
-	-	(-) - (-)	-	Variable has a stronger negative relationship with PA scores
-	-	(-) – (-)	+	Variable has a stronger negative relationship with MC scores
+	-	(+) – (-)	+	Variable has a positive relationship with PA scores and a negative relationship with MC scores
-	+	(-) – (+)	-	Variable has a negative relationship with PA scores and a positive relationship with MC scores

Results Tables: 4.10 – 4.22

Key

^{m} = marginal significance (> .05 and <= .09)
* = significant at the .05 level (p < .05)
** = significant at the .01 level (p < .01)

Note: Coefficients are reported with standard deviation provided in the parenthesis.

Ability Ranks 1 & 2 were merged as no significant difference was found in their coefficients for all three standardized tests.

Models in bold are the appropriate ones and their results are discussed in the text.

Table 4-10: Regression of ITBS Multiple Choice Test

Variable	Student Report		Teacher Report	
	Model 1	Model 1A	Model 1	Model 1A
Hands-on Science				
Teacher Scale			-.43 (.37)	.22 (.93)
Student Scale	.47** (.20)	1.28** (.31)		
Student Lever Dummy	-.33 (.34)	-.35 (.34)		
Student Classification Dummy	1.02^{m} (.58)	.89 (.58)		
Interaction of Ability Rank & Student Scale				
Scale X rank3		-.88^{m} (.48)		-.12 (.83)
Scale X rank4		-1.14* (.50)		-.61 (.96)
Scale X rank5		-1.50** (.44)		-1.93^{m} (1.00)
Female (male is reference)	-.87* (.41)	-.89* (.41)	-.74^{m} (.42)	-.77^{m} (.42)
Race/ethnicity (White is reference)				
Asian	-2.40** (.58)	-2.46** (.57)	-2.44** (.56)	-2.48** (.56)
Black	-4.99** (.61)	-5.00** (.62)	-5.19** (.62)	-5.16** (.63)
Hispanic	-2.77** (.56)	-2.66** (.55)	-3.05** (.59)	-3.00** (.59)
Classroom % Minority	-6.86** (2.06)	-6.46** (1.89)	-7.68** (2.14)	-7.75** (2.22)
Ability Rank (1&2 are reference)				
Rank 3	.83 (.90)	3.75* (1.67)	1.06 (.87)	1.39 (3.21)
Rank 4	3.67** (1.21)	7.52** (1.66)	3.85** (1.17)	5.81 (3.88)
Rank 5 (highest)	5.44** (1.32)	10.64** (1.66)	5.67** (1.28)	11.93** (3.87)
n	1231	1231	1231	1231
R^2	.2993	.3067	.2942	.3006

Table 4-11: Relationship of Student Hands-on Science Scale to ITBS Score by Ability Rank

Ability Rank	Coefficient on Scale	Coefficient on Interaction Term	Final Coefficient
Rank 1&2	1.28		1.28
Rank 3	1.28	-.88	.40
Rank 4	1.28	-1.14	.14
Rank 5	1.28	-1.50	-.22

Table 4-12: Regression of Lever Performance Assessment

Variable	**Student Report**		**Teacher Report**	
	Model 1	Model 1A	**Model 1**	Model 1A
Hands-on Science				
Teacher Scale			-.05 (.14)	.28 (.38)
Student Scale	.29** (.11)	.52** (.18)		
Student Lever Dummy	-.26 (.20)	-.25 (.20)		
Student Classification Dummy	.60^{m} (.32)	.55 (.33)		
Interaction of Ability Rank & Student Scale				
Scale X rank3		-.11 (.23)		-.17 (.43)
Scale X rank4		-.46^{m} (.24)		-.51 (.36)
Scale X rank5		-.42 (.25)		-.64 (.44)
Female (male is reference)	.07 (.19)	.07 (.19)	.13 (.20)	.11 (.20)
Race/ethnicity (White is reference)				
Asian	-.16 (.27)	-.17 (.27)	-.21 (.27)	-.21 (.26)
Black	-1.72** (.34)	-1.73** (.34)	-1.81** (.35)	-1.80** (.35)
Hispanic	-.97** (.28)	-.92** (.28)	-1.12)** (.30	-1.09** (.29)
% Minority	-3.07** (.82)	-3.02** (.81)	-3.40** (.87)	-3.44** (.90)
Ability Rank (1&2 are reference)				
Rank 3	.54 (.35)	.88 (.83)	.64 (.34)	1.17 (1.58)
Rank 4	1.58** (.47)	3.15** (.94)	1.66** (.46)	3.33* (1.37)
Rank 5	2.37** (.55)	3.81** (.99)	2.49** (.56)	4.54** (1.59)
n	1235	1235	1235	1235
R^2	.2419	.2459	.2298	.2334

Table 4-13: Regression of Classification Performance Assessment

Variable	**Student Report** **Model 1**	 Model 1A	**Teacher Report** **Model 1**	 Model 1A
Hands-on Science				
Teacher Scale			-.46 (.58)	-.17 (1.30)
Student Scale	.98** (.29)	1.13^{m} (.60)		
Student Lever Dummy	.11 (.60)	.11 (.60)		
Student Classification Dummy	2.57* (.97)	2.56* (.99)		
Interaction of Ability Rank & Student Scale				
Scale X rank3		-.66 (.93)		-.47 (1.21)
Scale X rank4		-.08 (.85)		.17 (1.25)
Scale X rank5		-.14 (.87)		-1.10 (1.54)
Female (male is reference)	1.23* (.55)	1.33* (.56)	1.53** (.57)	1.51* (.58)
Race/ethnicity (White is reference)				
Asian	-1.29 (1.15)	-1.30 (1.14)	-1.47 (1.14)	-1.46 (1.14)
Black	-6.72** (.98)	-6.73** (.99)	-7.07** (1.01)	-7.00** (1.02)
Hispanic	-5.01** (.97)	-5.02** (.97)	-5.63** (.97)	-5.61** (.98)
% Minority	-11.17** (2.66)	-11.14** (2.62)	-12.70** (2.60)	-12.67** (2.64)
Ability Rank (1&2 are reference)				
Rank 3	2.00* (.94)	4.31 (3.14)	2.41* (.95)	3.94 (4.18)
Rank 4	4.30** (1.42)	4.53 (2.94)	4.55** (1.41)	3.95 (4.32)
Rank 5	6.89** (1.60)	7.35* (3.03)	7.23** (1.63)	10.77* (4.87)
n	1224	1224	1224	1224
R^2	.2693	.2698	.2559	.2570

Table 4- 14: Significant Coefficients Represented as Standard Deviation Units of Test Scores

Variable	**ITBS** (Model 1A)	**Lever** (Model 1)	**Classification** (Model 1)
Hands-on Science			
Student Scale	.17	.08	.08
Student Classification Dummy			.21
Interaction of Ability Rank and Student Scale			
Scale X rank4	-.15		
Scale X rank5	-.20		
Female (male is reference)	-.12		.10
Race/ethnicity (White is reference)			
Asian	-.32		
Black	-.65	-.48	-.55
Hispanic	-.35	-.27	-.41
Classroom % Minority	-.84	-.85	-.91
Ability Rank (Ranks 1&2 are reference)			
Rank 3	.49		.16
Rank 4	.98	.44	.35
Rank 5 (highest)	1.39	.65	.56

Note: The standard deviations for the tests are as follows:

ITBS:	7.68
Lever:	3.62
Classification:	12.24

Table 4-15: Differential Relationship of Lever Performance Test Scores and ITBS Multiple Choice Test Scores

Variable	Coefficient for (Lever - ITBS)	
	Model 2	Model 2A
Hands-on Science		
Student Scale	.09 (.25)	-.20 (.39)
Student Lever Dummy	-.48 (.57)	-.48 (.57)
Student Classification Dummy	.13 (1.00)	.16 (.97)
Interaction of Ability Rank & Student Scale		
Scale X rank3		.61 (.61)
Scale X rank4		.25 (.49)
Scale X rank5		.49 (.64)
Female (male is reference)	1.21* (.58)	1.22* (.58)
Race/ethnicity (White is reference)		
Asian	2.87** (.70)	2.90** (.70)
Black	1.78* (.84)	1.79** (.85)
Hispanic	1.09 (.87)	1.08 (.87)
Classroom % Minority	.25 (1.85)	.16 (1.87)
Ability Rank (1&2 are reference)		
Rank 3	.73 (.61)	-1.38 (2.09)
Rank 4	-.02 (.89)	-.86 (1.49)
Rank 5 (highest)	-.02 (.91)	-1.73 (2.19)
n	1138	1138
R^2	.0223	.0232

Table 4-16: Differential Relationship of Classification Performance Test Scores and ITBS Multiple Choice Test Scores

Variable	**Coefficient for (Classification - ITBS)**	
	Model 2	**Model 2A**
Hands-on Science		
Student Scale	-.1 (.23)	-.88^{m} (.45)
Student Lever Dummy	.59 (.42)	.60 (.43)
Student Classification Dummy	.31 (.93)	.51 (.94)
Interaction of Ability Rank & Student Scale		
Scale X rank3		.67 (.77)
Scale X rank4		1.42* (.61)
Scale X rank5		1.99** (.55)
Female (male is reference)	2.04** (.67)	.2.06* (.66)
Race/ethnicity (White is reference)		
Asian	2.14** (.65)	2.22** (.65)
Black	.93 (.95)	.94 (.96)
Hispanic	-.37 (.79)	-.53 (.80)
Classroom % Minority	-.40 (1.48)	-.67 (1.45)
Ability Rank (1&2 are reference)		
Rank 3	.82 (1.00)	-1.33 (2.90)
Rank 4	-1.15 (1.07)	-5.94* (2.26)
Rank 5 (highest)	-1.27 (1.03)	-8.19** (2.02)
n	1129	1129
R^2	.0272	.0359

Table 4-17: Within and Between Variation in Hands-on Science and ITBS Test Scores

Variable	**Student Report**	
	Model 3	**Model 3A**
Hands-on Science		
Classroom Mean of Student Scale	-.20 (.26)	1.65** (.56)
Student Deviation from Classroom Mean of Scale	1.04** (.25)	1.16** (.23)
Interaction of Ability Rank & Classroom Mean of Hands-on Scale		
Mean X rank3		-1.57^{m} (.86)
Mean X rank4		-2.52** (.75)
Mean X rank5		-3.73** (.76)
Female (male is reference)	$-.84^{m}$ (.42)	-.90* (.39)
Race/ethnicity (White is reference)		
Asian	-2.20** (.54)	-2.18** (.52)
Black	-4.88** (.61)	-4.69** (.60)
Hispanic	-2.76** (.57)	-2.47** (.53)
Classroom % Minority	-7.88** (2.07)	-7.68** (2.06)
Ability Rank (1&2 are reference)		
Rank 3	.90 (.89)	6.18* (3.00)
Rank 4	3.67** (1.21)	12.31** (2.30)
Rank 5 (highest)	5.52** (1.30)	18.52** (2.09)
n	1231	1231
R^2	.3054	.3235

Table 4-18: Relationship of Classroom Mean of Student Scale with ITBS Score by Ability Rank

Ability Rank	Coefficient on Classroom Mean Scale	Coefficient on Classroom Mean Scale Interaction Term	Final Between- Class Coefficient
Rank 1&2	1.65		1.65
Rank 3	1.65	-1.57	.07
Rank 4	1.65	-2.52	-.88
Rank 5	1.65	-3.73	-2.09

Table 4-19: Within and Between Variation in Hands-on Science and Lever Test Scores

Variable	**Student Report**	
	Model 3	**Model 3A**
Hands-on Science		
Classroom Mean of Student Scale	.23^{m} (.15)	.87** (.32)
Student Deviation from Classroom Mean of Scale	.37** (.14)	.39** (.13)
Interaction of Ability Rank and Classroom Mean of Hands-on Scale		
Mean X rank3		-.63 (.41)
Mean X rank4		-.98* (.37)
Mean X rank5		-1.09* (.49)
Female (male is reference)	.10 (.19)	.08 (.19)
Race/ethnicity (White is reference)		
Asian	-.13 (.27)	-.12 (.27)
Black	-1.7** (.34)	-1.65** (.33)
Hispanic	-.97** (.29)	-.88** (.28)
Classroom % Minority	-3.28** (.83)	-3.19** (.83)
Ability Rank (1&2 are reference)		
Rank 3	.57* (.35)	2.72^{m} (1.43)
Rank 4	1.58** (.47)	4.97** (1.25)
Rank 5 (highest)	2.42** (.55)	6.18** (1.63)
n	1235	1235
R^2	.2393	.2480

Table 4-20: Relationship of Classroom Mean of Student Scale with Lever Score by Ability Rank

Ability Rank	Coefficient on Between-Class Scale	Coefficient on Between-Class Interaction Term	Final Between- Class Coefficient
Rank 1&2	.87		.87
Rank 3	.87		.87
Rank 4	.87	-.98	-.09
Rank 5	.87	-1.09	-.22

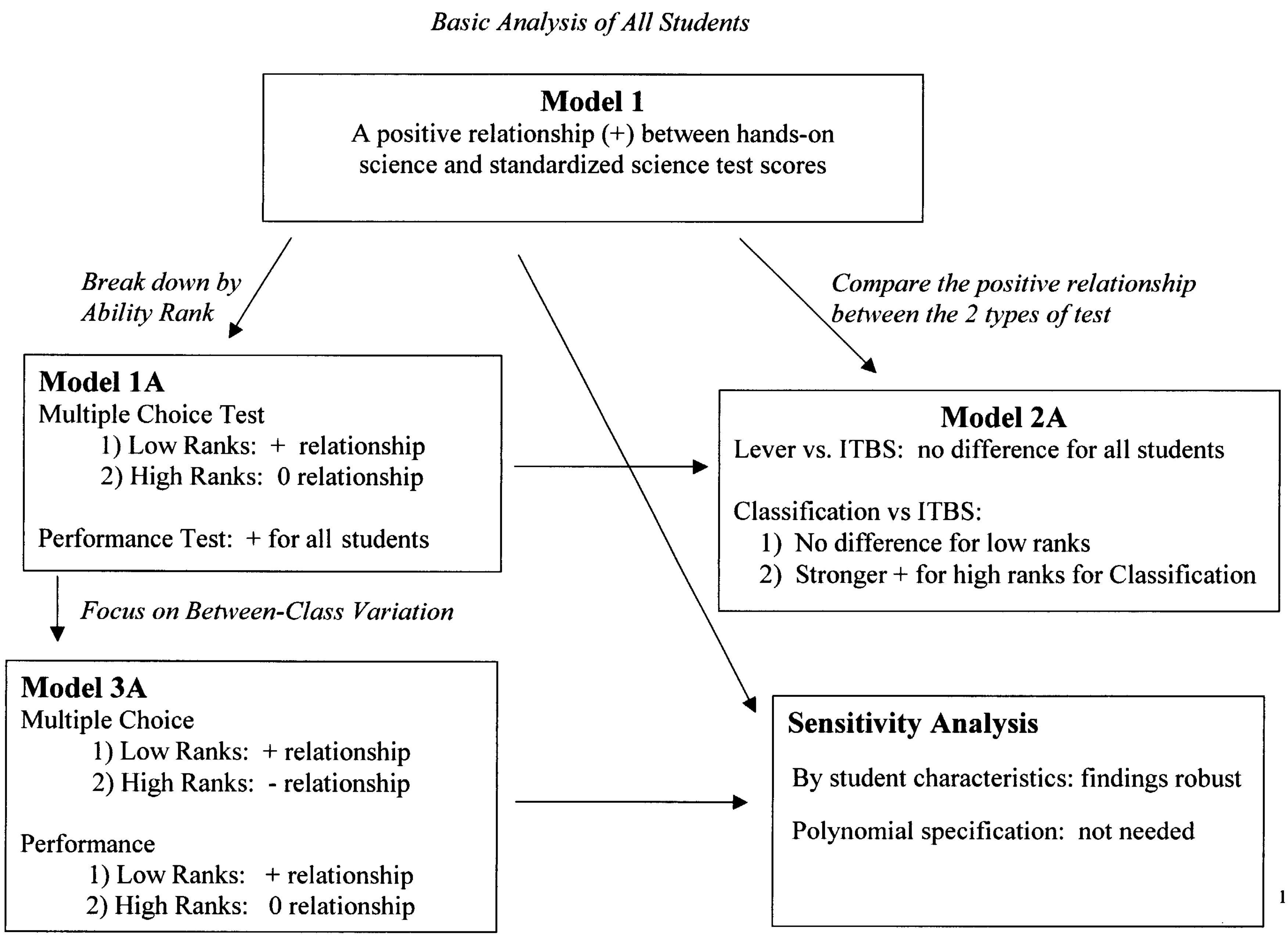
Table 4-21: Within and Between Variation in Hands-on Science and Classification Test Scores

Variable	**Student Report**	
	Model 3	**Model 3A**
Hands-on Science		
Classroom Mean of Student Scale	.43 (.54)	2.17* (.95)
Student Deviation from Classroom Mean of Scale	1.59** (.33)	1.65** (.32)
Interaction of Ability Rank & Classroom Mean of Hands-on Scale		
Mean X rank3		-1.68 (1.23)
Mean X rank4		-2.54* (1.13)
Mean X rank5		-3.11* (1.33)
Female (male is reference)	1.38* (.55)	1.34* (.56)
Race/ethnicity (White is reference)		
Asian	-1.10 (1.14)	-1.08 (1.11)
Black	-6.61** (.97)	-6.44** (.95)
Hispanic	-5.03** (.97)	-4.78** (.97)
Classroom % Minority	-12.51** (2.57)	-12.24** (2.56)
Ability Rank (1&2 are reference)		
Rank 3	2.10* (.96)	7.80^{m} (4.27)
Rank 4	4.23** (1.43)	12.95** (3.50)
Rank 5 (highest)	6.97** (1.61)	17.75** (3.86)
n	1224	1224
R^2	.2677	.2733

Table 4-22: Relationship of Classroom Mean of Student Scale with Classification Score by Ability Rank

Ability Rank	Coefficient on Between-Class Scale	Coefficient on Between-Class Interaction Term	Final Between- Class Coefficient
Rank 1&2	2.17		2.17
Rank 3	2.17		2.17
Rank 4	2.17	-2.54	-.37
Rank 5	2.17	-3.11	-.94

Figure 1: Analysis of RAND Data and Major Findings



Basic Analysis of All Students

Model 1
A positive relationship (+) between hands-on science and standardized science test scores

Break down by Ability Rank

Compare the positive relationship between the 2 types of test

Model 1A
Multiple Choice Test
1) Low Ranks: + relationship
2) High Ranks: 0 relationship

Performance Test: + for all students

Model 2A
Lever vs. ITBS: no difference for all students

Classification vs ITBS:
1) No difference for low ranks
2) Stronger + for high ranks for Classification

Focus on Between-Class Variation

Model 3A
Multiple Choice
1) Low Ranks: + relationship
2) High Ranks: - relationship

Performance
1) Low Ranks: + relationship
2) High Ranks: 0 relationship

Sensitivity Analysis

By student characteristics: findings robust

Polynomial specification: not needed

Appendix 4-1: Test of Use of Interaction Terms

(Model 1, 2 or 3 vs. Model 1A, 2A or 3A, respectively)

H_0: No difference between the two model (no need for interaction terms)
H_1: Difference between the two models

$$F_{(\text{dfho - dfh1, dfRSSH1})} = \frac{(RSS_{HO} - RSS_{H1}) / (df_{HO} - df_{H1})}{RSS_{H1} / df_{H1}}$$

$df_{HO} - df_{H1}$ = # of interaction terms
dfRSSH1 = degrees of freedom of H_1
RSS_{H1} / df_{H1} = MS Residual of H_1

1. ITBS

$$F_{(3,1214)} \text{ student} = \frac{(50746.57 -50214.41) / (1217 - 1214)}{41.36} = 4.29 \qquad p < .01$$

$$F_{(3,1216)} \text{ teacher} = \frac{(51116.44 - 50656.26) / (1219 -1216)}{41.66} = 3.68 \qquad p < .05$$

2. Lever

$$F_{(3,1218)} \text{ student} = \frac{(12280.21 - 12214.53) / (1221 -1218)}{10.03} = 2.18 \qquad p > .5$$

$$F_{(3,1220)} \text{ teacher} = \frac{(12475.78 - 12418.21) / (1223 - 1220)}{10.18} = 1.89 \qquad p > .5$$

3. Classification

$$F_{(3,1207)} \text{ student} = \frac{(133610.83- 133528.15) / (1210 - 1207)}{110.63} = .25 \qquad p > .10$$

$$F_{(3,1209)} \text{ teacher} = \frac{(136077.81 - 135871.11) / (1212-1209)}{112.38} = .61 \qquad p > .10$$

4. Difference Equation (Lever - MC)

$$F_{(3,1121)} \text{ student} = \frac{(95582.92 - 95504.50) / (1124 - 1121)}{85.20} = .31 \qquad p > .10$$

5. Difference Equation (Classification - MC)

$$F_{(3,1112)} \text{ student} = \frac{(96089.63 - 95232.17) / (1115 - 1112)}{85.64} = 3.34 \qquad p < .05$$

6. Within and Between Variation in Hands-on Science (ITBS)

$$F_{(3,1215)} \text{ student} = \frac{(50310.42 - 48999.86) / (1218 - 1215)}{40.33} = 10.83 \qquad p < .01$$

7. Within and Between Variation in Hands-on Science (Lever)

$$F_{(3,1219)}\text{ student} = \frac{\underline{(12321.77 - 12181.60) / (1222 - 1219)}}{9.99} = 4.68 \qquad p < .01$$

8. Within and Between Variation in Hands-on Science (Classification)

$$F_{(3,1208)}\text{ student} = \frac{\underline{(133908.27 - 132878.65) / (1211 - 1208)}}{110.00} = 3.12 \qquad p < .05$$

Note: F-tests are done on Models before Huber Correction is used.

Appendix 4-2: Checking the Robustness of the Results

Primarily, two models were used to analyze the RAND data. Model 1 (and its extension Model 1A which included interaction terms between hands-on science and Ability Rank) and Model 3A. The difference between them concerns the hands-on scale. Model 1 uses the hands-on science variable as reported by students. Model 3A decomposes that scale into: 1) the between-class difference in hands-on science, and 2) the within-class difference in hands-on science

Two types of extensions of the models were estimated. First, models were estimated to determine if the results changed for specific categories of students. Five specific categories of students were analyzed in this way:

1. Gender: female and male
2. Race/ethnicity: Asian, Black, Hispanic, White
3. Ability rank: 1 – 5 with 5 the highest
4. Within-class student agreement on reporting hands-on science (2 groups analyzed):
    a. Students from classes with low versus high variation in reports
    b. Students with low versus high deviation from their classroom mean
5. Students from classes with low versus high agreement with teacher reports of hands-on science (classes with high agreement were selected from classes found in 4a to have low variation in reports)

This first extension was done in two ways:

1. An interaction term between each of these categories and Hands-on Science was created. For Model 3A, an interaction was composed using between-class variation in hands-on science (interaction terms using within-class variation were not found significant). One at a time, these interaction terms were introduced into the original model which was then re-estimated.
2. Each model was estimated using data only from one specific category of students at a time.

The second extension addressed possible non-linearities through an introduction of a squared hands-on science term and, where appropriate, an introduction of interaction terms between the squared term and Ability Rank. The models were then estimated with these terms.

Results for Different Categories of Students

A. Estimation Using Interaction Terms

Table A4-1 reports the results from the separate estimation of Model 1A (for ITBS) and Model 1 (for Lever and Classification) when including the interaction terms between the student categories and hands-on science. The results from the original model (which also includes the interaction with ability rank) are given first. For the next five models, the table shows the interaction terms in bold. The table reports the coefficients (with the standard errors in parentheses) for the hands-on science variable, the interaction term to be tested, and, for Model 1A, the interaction between hands-on science and ability rank.

For the five extensions of Models 1 and 1A estimated[11], the coefficient for Hands-on Science (HO) remains positive as in the original model and near the magnitude of the original coefficient. It remains significant as well except in the case of the race interactions where it becomes marginal significant for Lever (p = .083) and non-significant for Classification (p = .094). The interaction terms are not significant. In a few cases they are marginal significant: Asian*HO for ITBS, CLATR*HO for Lever, and SLDFM*HO for Classification.

In the case of Model 1A (for ITBS), the interaction terms composed of hands-on science and ability rank remain negative and significant for the two highest ability ranks.

Table A4-2 reports the results from the estimation of Model 3A when including the interaction terms between the student categories and the classroom mean of the hands-on scale. These models include two sets of interaction terms (student category*classroom mean and ability rank*classroom mean).

For the five models estimated, the coefficient for the classroom mean remains positive and significant (except in the case of the race interaction model for Classification test scores where it is marginal significant). The majority of the introduced interaction terms were not found significant. Only the interaction term between hands-on science and students from classrooms reporting high agreement with teacher reports was significant (for Lever test scores). While this coefficient was negative, the coefficient for the classroom mean remained positive. Two other interaction terms had marginal significant coefficients (Asian*HO and Black*HO) for ITBS.

The five models with interaction terms also gave two other similar results to the original model. The interaction of hands-on science and ability rank was found negative and significant for the two highest ability ranks. The coefficient for student deviation from the classroom mean of the hands-on scale was found positive and significant.

[11] Note that Tables A4-1 through A4-3 compare the extended models to the original ones. They are not compared to one another and there is no comparison of models without interaction terms with models that have interaction terms.

In sum, our findings from this extension do not differ from those of our original model. Thus the findings from our original model are robust in regard to this extension.

B. Estimation By Individual Student Category

Tables A4-3 & A4-4 report the results from Models 1, 1A, and 3A when estimated separately using data from specific student groups within the categories of gender, race/ethnicity, and ability rank. For example, each table lists a horizontal panel titled Gender. In this panel are the results when the models are estimated separately for all females and for all males.

The results from Models 1 & 1A shown in Table A4-3 are similar to those from the original model which are given at the top of the table. The coefficients for hands-on science (HO) are positive as in the original model. That not all of them remain significant may in part be due to the smaller sample sizes. For Model 1A (ITBS) the interaction between hands-on science and ability rank remains negative and significant for the top two ranks. The only exception to this is when the model was run for Hispanics, these coefficients were positive (and non-significant). This may be due the lack of Hispanics in the top two ability ranks.

The results from Model 3A (Table A4-4) regarding between-class variation in hands-on science are similar to the original model. The coefficient for the classroom mean of the hands-on scale remains positive (and often significant) while the interaction of the classroom mean and ability rank remains negative for the top two ability ranks. Exceptions to the latter, occur with Classification test scores for Black, Hispanic and White for which the interaction of HO*rank5 is positive. These may be exhibiting less precise estimates based on the small number of observations in these cells, at least for Black and Hispanic.

I then created two categories of students based on their agreement with their classmates reports of hands-on science. These categories were created in two ways: 1) choosing all students from classes that exhibited high or low agreement, and 2) choosing students who exhibited high or low agreement with their classroom mean. As an additional step, I created two categories based on student agreement with teacher reports: classrooms that had high student agreement were matched to teacher reports to determine if they also had high student-teacher agreement. As noted above, these variables were used to make interaction terms by combining them with hands-on science. Here, the Models 1, 1A and 3A were estimated separately for high versus low categorized students (for each of these three categories).

The analyses of these three categories provided similar results. In order to save space, I will only report the one with the least similar results to the original model: students from classes with a high standard deviation of within-class variation in hands-on reports versus those from classes with a low standard deviation.

Table A4-5 gives the results from Model 1A (for ITBS). The results are very similar to the original model. Both categories of students have positive significant (or marginal significant) coefficients on Hands-on Science and negative coefficients on the interaction terms (hands-on science*ability rank) for the highest ability ranks (significant for students from classes with a low standard deviation and marginal significant for students from high deviation classes).

Table A4-6 gives the results from Model 1 (for Lever and Classification). Students with a high deviation give results similar to the original model while students with a low deviation having a coefficient on Hands-on Science that is much smaller and non-significant. This might occur because the hands-on science variable contains both between-class and within-class variation and by definition the low deviation classes have lower-within class variation. Therefore, they have less variation in the independent variable of interest.

Table A4-7 gives the results from Model 3A. They show that students with a high deviation give results similar to those from the original model while students with a low deviation give results of the same sign as the original model but of lower magnitude and significance level. The lower significance level may be due to the smaller sample size (511 students).

The results when using the category of students with low versus high agreement with their classroom mean (not shown here) has a sample size of 841 for high agreement and the results are significant for this group (but still of lower magnitude than the group composed of students with low agreement) for both ITBS and Lever when estimating Model 3A.

Results for Quadratic Models

The squared hands-on science term was introduced in two ways into the original Models 1, 1A and 3A. First, the models were estimated with the squared term. Second, the models (except Model 1) were estimated using the squared term and its interaction with the ability rank terms.

Tables A4-8 and A4-9 show the results for Model 1A (ITBS) and Model 1 (Lever and Classification) respectively. For ITBS, neither the squared term is significant nor are the interaction terms. For Classification, the squared term is not significant. For Lever, the squared term is significant but with its introduction, the main effect becomes negative and non-significant. None of these results support the use of a polynomial specification for Models 1 and 1A.

Tables A4-10 and A4-11 show the results of Model 3A for the three test scores. For ITBS, neither the squared term is significant nor are its interaction terms. As a result of introducing them, the coefficient on the hands-on scale becomes marginal significant (but remains positive) while the interaction terms of hands-on science and ability rank remain negative and significant. For Lever, the squared term is marginal significant and its interaction with the highest rank is positive and significant. With its introduction, the coefficient on the hands-on scale becomes marginal significant (but remains positive) while the interaction of hands-on science and ability rank remain negative for the higher ability ranks and significant for the top rank. For Classification, the squared term is negative and significant and one of its interaction terms (with ability rank 4) is positive and significant. The coefficient on the hands-on scale remains positive and significant and the interaction terms of hands-on science and ability remain negative for the higher ability ranks and significant or marginal significant for the top two ranks.

For all three types of test, we see a similar pattern in the sign of the coefficients. The main hands-on scale is positive, the interaction terms of hands-on science and ability rank is negative for the higher ability ranks, the squared hands-on scale is negative and the interaction terms composed of the squared term and ability rank is positive. The expected ability rank effect (the coefficient on the interaction term offsets the coefficient on the respective hands-on science term) occurs in all cases.

Only in the case of Classification, do we have evidence (a significant squared term and interaction term) that a polynomial specification may be correct and that after some point too much Hands-on Science has a negative relationship with Classification test scores. When this model was run separately for each ability rank (estimating Model 3 for Classification), it was found that the significance of the squared term only held for Ability Ranks 1, 2 and 3. Therefore, we have some evidence that hands-on science's positive relationship to Classification test scores stops at some level of hands-on science (in this case, this occurs at 3.44 on the hands-on scale of 1 – 5) for lower ability student.

Summary

Overall, the results from the further extensions of Models 1, 1A and 3A show the robustness of the results from the original models. Concerning the extensions containing interaction terms for different groups, we found that the coefficient on hands-on science remained positive and significant (with two exceptions in Models 1 & 1A and one exception in Model 3A), that the interaction of hands-on science and ability rank remained negative and significant for the higher ability ranks (in Models 1A and 3A) and that only in one case was the newly introduced interaction term significant. When we checked these results by estimating the original models individually for each category of student (the same categories are those that made up the introduced interaction terms), the same results were on the whole confirmed. The coefficient on the hands-on scale remained positive (except in one case) and often significant (though sometimes not due in part to a reduced sample size) for Models 1, 1A and 3A. The interaction of Hands-on Science and ability rank remained negative (with one exception in Model 1A and three in Model 3A.

We examined a further extension using a squared hands-on science term and its interactions with ability rank and did not find evidence for using such a polynomial specification in the case of ITBS and Lever. However, we did find some evidence for its use in the case of Classification and further examination showed that the positive relationship of hands-on science and Classification test score might be limited within a specific range of hands-on science for lower ability rank students. We also confirmed the finding in the original models, that the positive relationship of hands-on science to test score is reduced for higher ability ranks.

In sum, the original model specifications appear to be robust. The possible importance of a polynomial specification for the Classification test should be noted along with its implications.

Table A4-1: Models 1 and 1A Using Interaction Terms

Specification and Variables	ITBS Model 1A	Lever Model 1	Classification Model 1
Original Model			
HO Scale	1.28** (.31)	.29** (.11)	.98** (.29)
Rank3*HO	-.88^{m} (.48)		
Rank4*HO	-1.14* (.50)		
Rank5*HO	-1.50** (.44)		
Gender Interaction (male as reference)			
HO scale	1.47** (.34)	.36** (.11)	1.26** (.33)
Female*HO scale	-.43 (.29)	-.14 (.17)	-.54 (.39)
Rank3*HO	-.86^{m} (.48)		
Rank4*HO	-1.11* (.50)		
Rank5*HO	-1.43** (.45)		
Race/Ethnicity Interaction (white as reference)			
HO scale	1.55** (.39)	.28^{m} (.16)	.85 (.49)
Asian*HO scale	-.67^{m} (.38)	-.08 (.23)	.59 (.81)
Black*HO scale	-.31 (.52)	-.03 (.28)	-.25 (.71)
Hispanic*HO scale	-.29 (.40)	.13 (.19)	.15 (.59)
Rank3*HO	-.92^{m} (.48)		
Rank4*HO	-1.17* (.51)		
Rank5*HO	-1.49** (.48)		
Interaction with Students from classrooms with overall high deviation from mean (CLDFM = 0 or 1)			
HO scale	1.34** (.40)	.25 * (.11)	1.06** (.31)
CLDFM*HO scale	-.07 (.19)	.06 (.08)	-.13 (.21)
Rank3*HO	-.89^{m} (.49)		
Rank4*HO	-1.16* (.51)		
Rank5*HO	-1.52** (.46)		
Interaction with Students with high deviations from their classroom mean (SLDFM = 0 or 1)			
HO scale	1.25** (.30)	.26* (.10)	.85** (.29)
SLDFM*HO scale	.08 (.09)	.07 (.05)	.33^{m} (.18)
Rank3*HO	-.88^{m} (.48)		
Rank4*HO	-1.14* (.49)		
Rank5*HO	-1.50** (.44)		
Interaction with Students from classrooms with high agreement with teacher reports and low student deviation from mean (CLATR = 0 or 1)			
HO scale	1.31** (.31)	.33** (.11)	.99** (.31)
CLATR*HO scale	-.13 (.11)	-.15^{m} (.08)	-.05 (.21)
Rank3*HO	-.87^{m} (.49)		
Rank4*HO	-1.13* (.50)		
Rank5*HO	-1.50** (.44)		

Table A4-2: Model 3A Using Interaction Terms

Specification	ITBS	Lever	Classification
Original Model			
Between-Class HO Scale	1.65** (.56)	.87** (.32)	2.17* (.95)
Rank3*HO	-1.57^{m} (.86)	-.63 (.41)	-1.68 (1.23)
Rank4*HO	-2.52** (.75)	-.98* (.37)	-2.54* (1.13)
Rank5*HO	-3.73** (.76)	-1.09* (.49)	-3.11* (1.33)
Within-Class HO Scale	1.11** (.23)	.39** (.13)	1.65** (.32)
Gender Interaction (male as reference)			
Between-Class HO scale	1.73** (.54)	.92* (.35)	1.98* (.91)
Female*HO scale	-.21 (.45)	-.12 (.26)	.45 (.70)
Rank3*HO	-1.56^{m} (.88)	-.63 (.41)	-1.73 (1.23)
Rank4*HO	-2.49** (.78)	-.97* (.36)	-2.60* (1.16)
Rank5*HO	-3.66** (.82)	-1.05* (.50)	-3.25* (1.39)
Within-Class HO scale	1.11** (.23)	.39** (.13)	1.65** (.32)
Race/Ethnicity Interaction (white as reference)			
Between-Class HO scale	2.40** (.79)	1.11* (.44)	2.52^{m} (1.40)
Asian*HO scale	-1.12^{m} (.59)	.13 (.29)	-.12 (1.10)
Black*HO scale	-1.62^{m} (.88)	-.63 (.41)	-.38 (1.30)
Hispanic*HO scale	-.68 (.69)	-.29 (.36)	-.60 (1.11)
Rank3*HO	-1.77^{m} (.90)	-.72 (.43)	-1.76 (1.26)
Rank4*HO	-2.70** (.79)	-1.09** (.41)	-2.71* (1.31)
Rank5*HO	-3.98** (.91)	-1.31* (.58)	-3.36* (1.59)
Within-Class HO scale	1.07** (.24)	.38** (.13)	1.64** (.31)
Interaction with Students from classrooms with overall high deviation from mean (CLDFM = 0 or 1)			
Between-Class HO scale	1.77** (.65)	.85* (.35)	2.38* (1.00)
CLDFM*HO scale	-.13 (.19)	.02 (.08)	-.23 (.22)
Rank3*HO	-1.65^{m} (.89)	-.62 (.43)	-1.81 (1.19)
Rank4*HO	-2.60** (.80)	-.97* (.39)	-2.68* (1.17)
Rank5*HO	-3.81** (.83)	-1.07 (.51)	-3.25* (1.36)
Within-Class HO scale	1.11** (.23)	.39** (.13)	1.65** (.32)
Interaction with Students with high deviations from their classroom mean (SLDFM = 0 or 1)			
Between-Class HO scale	1.64** (.57)	.85* (.32)	2.08* (.97)
SLDFM*HO scale	.01 (.09)	.06 (.05)	.25 (.20)
Rank3*HO	-1.57^{m} (.86)	-.63 (.41)	-1.66 (1.22)
Rank4*HO	-2.52** (.75)	-.97* (.37)	-2.48* (1.44)
Rank5*HO	-3.73** (.76)	-1.08* (.49)	-3.08* (1.32)
Within-Class HO scale	1.11** (.24)	.41** (.14)	1.71** (.33)

table A4-2 continued

Table A4-2

Specification	ITBS	Lever	Classification
Interaction with Students from classrooms with high agreement with teacher reports and low student deviation from mean (CLATR = 0 or 1)			
Between-Class HO scale	1.67** (.55)	.93** (.30)	2.21* (.97)
CLATR*HO scale	-.10 (.11)	-.16* (.07)	-.11 (.21)
Rank3*HO	-1.56^{m} (.86)	-.64 (.41)	-1.68 (1.22)
Rank4*HO	-2.51** (.75)	-.98** (.36)	-2.53* (1.13)
Rank5*HO	-3.72** (.75)	-1.08* (.48)	-3.11* (1.31)
Within-Class HO scale	1.11** (.23)	.39** (.13)	1.65** (.32)

Table A4-3: Models 1 & 1A Estimated Separately By Gender, Racial/Ethnic and Ability Rank Groups

Specification & Variables	ITBS Model 1A	Lever Model 1	Classification Model 1
Original HO	1.28** (.31)	.29** (.11)	.98** (.29)
HO*Rank 3	-.88^{m} (.48)		
HO*Rank 4	-1.14* (.50)		
HO*Rank 5	-1.50** (.44)		
Gender			
Female HO	.89* (.39)	.17 (.17)	.71* (.36)
HO*Rank 3	-.57 (.56)		
HO*Rank 4	-1.00^{m} (.59)		
HO*Rank 5	-1.20* (.60)		
Male HO	1.50** (.44)	.38** (.11)	1.20** (.34)
HO*Rank 3	-1.20^{m} (.66)		
HO*Rank 4	-.95 (.73)		
HO*Rank 5	-1.68* (.72)		
Race/ethnicity			
Asian HO	1.34 (.84)	.18 (.16)	1.41* (.57)
HO*Rank 3	-2.08^{m} (1.18)		
HO*Rank 4	-1.18 (1.32)		
HO*Rank 5	-2.29* (.94)		
Black HO	.82 (.83)	.22 (.28)	.46 (.63)
HO*Rank 3	1.95 (1.14)		
HO*Rank 4	-1.59 (1.39)		
HO*Rank 5	-1.54 (2.51)		
Hispanic HO	.55 (.42)	.25 (.17)	.79 (.48)
HO*Rank 3	-.45 (.69)		
HO*Rank 4	.29 (.92)		
HO*Rank 5	.71 (1.08)		
White HO	2.25** (.68)	.32* (.16)	.76 (.49)
HO*Rank 3	-2.09* (.87)		
HO*Rank 4	-1.94* (.80)		
HO*Rank 5	-2.29** (.76)		
Ability Rank	Requires Model 1		
Ranks 1 & 2 HO	1.11** (.28)	.39* (.19)	.85 (.61)
Rank3 HO	.61 (.41)	.51* (.17)	.93 (.67)
Rank4 HO	.18 (.32)	.09 (.18)	.98^{m} (.53)
Rank 5 HO	-.29 (.34)	.10 (.15)	.79 (.53)

Table A4-4: Model 3A Estimated Separately By Gender, Racial/Ethnic and Ability Rank Groups

Specification & Variables	ITBS	Lever	Classification
Original Between HO	1.65** (.56)	.87** (.32)	2.17* (.95)
HO*Rank 3	-1.57^{m} (.86)	-.63 (.41)	-1.68 (1.23)
HO*Rank 4	-2.52** (.75)	-.98* (.37)	-2.54* (1.13)
HO*Rank 5	-3.73** (.76)	-1.09* (.49)	-3.11* (1.33)
Gender			
Female Between HO	1.02 (.86)	.69 (.42)	1.61 (1.70)
HO*Rank 3	-1.33 (1.05)	-.52 (.50)	-.60 (2.14)
HO*Rank 4	-1.58 (.98)	-.96^{m} (.49)	-1.82 (1.87)
HO*Rank 5	-2.94* (1.14)	-.86 (.52)	-1.36 (1.74)
Male Between HO	2.04** (.69)	1.04** (.33)	2.68** (.92)
HO*Rank 3	-1.78^{m} (1.01)	-.79 (.46)	-2.88 (1.69)
HO*Rank 4	-3.19** (1.01)	-.88* (.40)	-3.26* (1.49)
HO*Rank 5	-4.55** (1.09)	-1.49* (.75)	-6.19** (2.24)
Race/ethnicity			
Asian Between HO	3.41** (1.25)	1.31* (.60)	6.62** (1.05)
HO*Rank 3	-5.51** (1.99)	-1.00 (1.20)	-4.50 (3.65)
HO*Rank 4	-5.03** (1.43)	-1.05 (.78)	-9.74** (2.47)
HO*Rank 5	-6.43** (1.42)	-1.51^{m} (.79)	-8.12** (1.52)
Black Between HO	.65 (.95)	.03 (.63)	2.54 (2.14)
HO*Rank 3	2.84 (2.17)	1.69 (1.09)	2.47 (4.07)
HO*Rank 4	-3.60* (1.36)	-1.29 (.84)	-5.45 (3.54)
HO*Rank 5	-5.12^{m} (2.85)	-.09 (1.11)	8.10^{m} (4.24)
Hispanic Between HO	1.61** (.57)	.35 (.35)	1.21 (1.06)
HO*Rank 3	-2.42* (.94)	-.84 (.53)	-1.80 (1.62)
HO*Rank 4	-2.02 (1.39)	.19 (.58)	-1.53 (1.60)
HO*Rank 5	-3.43* (1.71)	-1.07 (1.18)	4.05 (4.16)
White Between HO	1.85 (1.36)	1.38* (.54)	-.64 (1.99)
HO*Rank 3	-.86 (1.56)	-.70 (.56)	1.74 (1.99)
HO*Rank 4	-1.92 (1.38)	-1.38* (.62)	1.96 (2.20)
HO*Rank 5	-2.97* (1.44)	-1.45* (.61)	.15 (2.30)
	Model 3	Model 3	Model 3
Ability Rank Between HO			
Ranks 1 & 2	1.37* (.64)	.72* (.36)	1.63 (1.00)
Rank 3	.34 (.63)	.36 (.24)	1.04 (1.07)
Rank 4	-.82^{m} (.42)	-.16 (.23)	-.62 (.89)
Rank 5	-1.98** (.58)	-.20 (.28)	-.99 (.72)

Table A4-5: Model 1A (ITBS) Estimated Separately for Students from Low Deviation Classes vs. those from High Deviation Classes

Variable	Students from All Classes	Students from Low Deviation Classes CLDFM = 0	Students from High Deviation Classes CLDFM = 1
Hands-on Scale	1.28** (.31)	1.30^{m} (.69)	1.17** (.33)
HO*Rank 3	-.88^{m} (.48)	-1.01 (.94)	-.89 (.59)
HO*Rank 4	-1.14* (.50)	-1.34 (1.00)	-1.03^{m} (.57)
HO*Rank 5	-1.50** (.44)	-2.48* (.93)	-.91^{m} (.50)
n	1231	517	714

Table A4-6: Model 1 (Lever & Classification) Estimated Separately for Students from Low Deviation Classes vs. those from High Deviation Classes

	LEVER				CLASSIFICATION		
Variable	Students from All Classes	Students from Low Deviation Classes CLDFM = 0	Students from High Deviation Classes CLDFM = 1		Students from All Classes	Students from Low Deviation Classes CLDFM = 0	Students from High Deviation Classes CLDFM = 1
Hands-on Scale	.29** (.11)	-.01 (.13)	.40* (.15)		.98** (.29)	.43 (.49)	1.21** (.38)
n	1235	519	716		1224	511	713

Table A4-7: Model 3A Estimated Separately for Students from Low Deviation Classes vs. those from High Deviation Classes

Test	Variable	Students from All Classes	Students from Low Deviation Classes CLDFM = 0	Students from High Deviation Classes CLDFM = 1
ITBS				
	Between-Class Hands-on Scale	1.65** (.56)	1.14 (.93)	2.32** (.77)
	HO*Rank 3	-1.57^{m} (.86)	-1.14 (1.38)	-2.92* (1.27)
	HO*Rank 4	-2.52** (.75)	-1.78 (1.16)	-4.48** (1.13)
	HO*Rank 5	-3.73** (.76)	-4.27** (1.13)	-3.69* (1.39)
	Within Class Hands-on Scale	1.11** (.23)	1.26** (.42)	1.03** (.28)
	n	1231	517	714
Lever				
	Between-Class Hands-on Scale	.87** (.32)	.45 (.44)	1.34** (.39)
	HO*Rank 3	-.63 (.41)	-.64 (.63)	-.87* (.36)
	HO*Rank 4	-.98* (.37)	-.81 (.48)	-1.29* (.60)
	HO*Rank 5	-1.09* (.49)	-.83 (.56)	-1.72* (.82)
	Within Class Hands-on Scale	.39** (.13)	.17 (.23)	.46** (.16)
	n	1235	519	716
Classification				
	Between-Class Hands-on Scale	2.17* (.95)	1.89 (1.32)	2.24 (1.67)
	HO*Rank 3	-1.68 (1.23)	-2.71 (1.74)	-.74 (1.69)
	HO*Rank 4	-2.54* (1.13)	-2.49^{m} (1.29)	-3.04 (2.23)
	HO*Rank 5	-3.11* (1.33)	-3.28^{m} (1.62)	-3.19 (2.57)
	Within Class Hands-on Scale	1.65** (.32)	1.40* (.64)	1.70** (.38)
	n	1224	511	713

Table A4-8: Model 1A (ITBS) With Quadratic

Variable	Original Model	With Squared Term	With Squared Term and Interactions
Hands-on Scale	1.28** (.31)	.04 (1.14)	1.34 (2.33)
HO*Rank 3	-.88^{m} (.48)	-.94^{m} (.49)	-.39 (.42)
HO*Rank 4	-1.14* (.50)	-1.15* (.48)	-3.97 (.45)
HO*Rank 5	-1.50** (.44)	-1.54** (.44)	-4.24 (.44)
Quadratic HO Scale		.19 (.17)	-.01 (.36)
Quad*Rank3			-.07 (.42)
Quad*Rank4			.44 (.45)
Quad*Rank5			.41 (.44)
R^2	.3067	.3077	.3090

Table A4-9: Model 1 (Lever & Classification) With Quadratic

	Lever		Classification	
Variable	Original Model	With Squared Term	Original Model	With Squared Term
Hands-on Scale	.29** (.11)	-.86 (.53)	.98** (.29)	-.37 (1.92)
Quadratic HO Scale		.18* (.08)		.21 (.30)
R^2	.2149	.2457	.2693	.2698

Table A4-10: Model 3A (ITBS) With Quadratic

Variable	Original Model	With Squared Term	With Squared Term and Interactions
HO Between-Class Scale	1.65** (.56)	4.23 (3.85)	13.29^{m} (7.61)
HO*Rank 3	-1.57^{m} (.86)	-1.43 (.90)	-13.85^{m} (7.14)
HO*Rank 4	-2.52** (.75)	-2.41** (.77)	-17.12* (7.91)
HO*Rank 5	-3.73** (.76)	-3.59** (.78)	-16.19* (7.30)
Quadratic HO Scale		-.42 (.62)	-1.87 (1.2)
Quad*Rank 3			1.97^{m} (1.10)
Quad*Rank 4			2.33^{m} (1.32)
Quad*Rank 5			2.00^{m} (1.15)
Within-Class Scale	1.11** (.23)	1.10** (.23)	1.13** (.23)
R^2	.3235	.3243	.3297

Table A4-11: Model 3A (Lever & Classification) With Quadratic

	Lever			Classification		
Variable	Original Model	With Squared Term	With Squared Term and Interactions	Original Model	With Squared Term	With Squared Term and Interactions
Between-Class HO Scale	.87** (.32)	2.25 (1.63)	6.11^{m} (3.11)	2.17* (.95)	10.41^{m} (5.33)	22.48* (8.46)
HO*Rank 3	-.63 (.41)	-.58 (.44)	-5.50 (3.52)	-1.68 (1.23)	-1.26 (1.25)	-5.89 (9.73)
HO*Rank 4	-.98* (.37)	-.93* (.39)	-5.51 (3.39)	-2.54* (1.13)	-2.15^{m} (1.20)	-28.77** (9.42)
HO*Rank 5	-1.09* (.49)	-1.02* (.51)	-9.52* (3.69)	-3.11* (1.33)	-2.69^{m} (1.40)	-20.90^{m} (12.20)
Quadratic HO Scale		-.22 (.26)	-.84^{m} (.48)		-1.37 (.87)	-3.27* (1.37)
Quad*Rank 3			.78 (.54)			.79 (1.42)
Quad*Rank 4			.73 (.51)			4.18* (1.48)
Quad*Rank 5			1.33* (.58)			2.88 (1.97)
Within-Class Scale	.39** (.13)	.39** (.13)	.41** (.13)	1.65** (.32)	1.64** (.32)	1.68** (.32)
R^2	.2480	.2491	.2547	.2733	.2769	.2833

Chapter 5: Analysis of The National Education Longitudinal Study (NELS:88)

Introduction

The National Educational Longitudinal Survey (NELS:88) followed students from 8th grade in 1988 to 10th grade in 1990 to 12th grade in 1992. Our analysis of the three waves of NELS data allows us to answer some questions raised by the analysis of the RAND data and to extend them. Will the NELS data continue to show a positive relationship between student reported hands-on science and multiple choice test, when we include additional covariates that may be related to both? If so, will we continue to see a differential relationship between hands-on science and test score due to student ability? Will we continue to see a lack of relationship when we use teacher reported data or was the small teacher sample size in the RAND data the cause of this finding? Also, will our RAND results hold for the two higher grades (10th and 12th) included in the NELS?

NELS began in 1988 with a survey of 8th grade students, their teachers, parents and schools. A cross-sectional analysis of the base year data can be compared to our analysis of the RAND 8th grade data. Our analysis of the RAND data found a positive relationship between hands-on work and multiple choice test scores for students of lower ability and no relationship for students of middle or higher ability when using student reports of hands-on science. No relationship was found when using teacher reports of hands-on science.

NELS allows us to further test Hypotheses 1 and 3 in several ways. NELS is a nationally representative sample which allows inferences to be made to the whole population. Second, it contains additional variables that may be correlated with test scores. If the results from the analysis of the NELS data show a positive relationship between test score and hands-on science scale (Hypothesis 1) while using additional controls, we will have greater confidence in the robustness of the finding as it is supported by two different data sets, the latter having a full set of control variables. We will also check whether the addition of covariates supports or explains away the important role of student ability in this relationship that we found in the RAND data (Hypothesis 3). Third, in the last chapter we postulated that the finding of a lack of a relationship between test score and the teacher reported hands-on scale may have been due to the small number of teachers involved in the RAND study. If the analysis of the NELS data replicates this finding, such a postulate cannot be supported. Conversely, if we do find a positive relationship, then the postulate will be reasonable. NELS also provides the opportunity for testing the relationship between hands-on science and test score for the students in grades 10 and 12 to determine if the findings for the 8th grade are consistent across grades or if they differ, possibly due to the increasing ability of more mature students to comprehend abstract ideas without the use of concrete examples.

This chapter is organized in the following fashion. We first discuss the NELS sample and identify the measures we will use in our analysis. Next we provide descriptive statistics of the sample. The model is then described, the results are discussed, and a set of sensitivity analyses is reported. Where applicable, we make comparisons with findings from the RAND data.

I. The Sample

In 1988, the National Education Longitudinal Study (NELS:88) surveyed 25,000 8th graders in 1000 public and private schools across the United States and resurveyed these students in 1990 (10th grade) and 1992 (12th grade). Using a two-stage stratified clustered sample design the sample was nationally representative in 1988 and was freshened (new students added) in both 1990 and 1992 so that it remained nationally representative in both follow-ups. The two-stage design first sampled schools in 1988 from the 39,000 schools that have an 8th grade. Within the selected schools, students were sampled at an average of 23 per school. The 1990 follow-up included all students who were in the base year sample and went to a school with ten or more other base year students. The remaining students from the base year were probability sampled and freshened students were added. In 1992, an attempt was made to survey all students who were in the first wave. In addition, freshened students were added (NCES 1994).

Measures of student achievement, level of hands-on science in the classroom, student background, and school characteristics were collected in each year through student testing, teacher surveys, parent surveys, student surveys, and school surveys. In 1988, two teachers of different subjects were to be surveyed for each student. The combination of teachers was to include one English or history teacher and one math or science teacher. In 1990, there was an attempt to do a similar selection of teachers but if the student was not taking the chosen subject at the time of the survey another teacher of the student was surveyed. This allowed two teachers of any combination of the four subjects to be surveyed per student. In 1992, only one teacher, math or science, was surveyed per student. If a student was not in a math or science course, there is no teacher survey data for him/her.

Due to missing data, the size of the sub-samples available for our analysis is substantially smaller than the total sample. First, not all students have a science test score. Second, students may have not answered the hands-on science items on the student survey in grade 10 or 12 (no hands-on science questions were asked of students in grade 8). Third, a larger number of students may lack a report of a teacher hands-on science scale because only a sub-sample of the students had their science teacher surveyed[1]. Out of the sample of 25,000 8th graders, 23,316 have test scores, and 10,221 have a teacher hands-on science scale. For 10th grade, out of 18,100 students, 16,500 have test scores, 15,500 have a student scale and 5,400 have a teacher scale. For 12th grade, out of 17,000 students, 12,700 have scores, 11,200 have a student scale and 3,200 have a teacher scale.

We want to check that the characteristics of the sub-samples available for our analysis do not greatly differ from those of the full NELS sample by examining whether there are significant differences in the descriptive statistics. In Appendix tables A5-1 to A5-3, we compare the descriptive statistics of our covariates in the full NELS sample versus the sub-samples that contain a test score, the student hands-on scale, and the teacher hands-on scale. Overall, these tables show that the distribution of such variables

[1] The lack of a teacher survey for a student may be due to one of three situations:

1. The student is in a science class and his/her science teacher was not picked to be surveyed
2. The student is in a science class, his/her science teacher was picked to be surveyed but did not return the survey
3. The student was not in a science class that year so there was no science teacher to be surveyed.

In all three cases, the student would not be included in the analysis of the teacher hands-on scale. The student would be included in the analysis of the student hands-on scale, in part because the student items as gender, race/ethnicity, ability rank, percent minority and the school variables are consistent between the full sample and the sub-samples we use in our analysis. However, there are changes in several variables that warrant cautions.

We see a drop in the percent of minority students (primarily Hispanic), a drop in the percent of the lowest SES quartile, and a drop in the percentage of the lower ability rank students in the subsamples. This drop is greatest in the 12th grade when looking at the sub-sample available for the teacher scale for which a large number of teacher responses are lacking and smallest in the 8th grade for which we have a greater number of teacher responses. For example, in 8th grade the lowest SES quartile makes up 24.2% of the full sample, 23.9% of the sub-sample with a science test score and 22.3% of the sub-sample with a teacher report (Table A5-1). In 12th grade, these percentages are 19%, 17.2% and 13.3% respectively (Table A5-3). Our analyzed data may then slightly under-represent lower achieving students.

ask for the level of hands-on science done in the last science course taken. A further discussion will be provided regarding differences in findings among students taking science versus those not taking science.

II. Measures

Three types of measures were obtained from the NELS materials. Using responses to the teacher and student surveys, we created scales to measure the quantity of hands-on science in the classroom. Measures of student achievement are based on student scores on the science multiple choice test. In addition, covariates that may be related to test score were obtained from several of the NELS instruments.

Science Test Score

Students within the same grade took the same science test. The test was similar between grades with some substitution of new questions to reduce ceiling effects. Each grade's test contained 25 multiple choice questions and was to be completed in 20 minutes. The 10th grade test contained 18 questions from the 8th grade test and 7 new questions. The 12th grade test contained 19 questions from the 10th grade test (13 of which were also on the 8th grade test), and 6 new items. In total, there was a pool of 38 science items. To reduce ceiling effects, the new questions were either more difficult or addressed content knowledge normally taught in more advanced classes.

In order to measure gains in test scores across grades, these tests are calibrated on the same scale using Item Response Theory (NCES 1995). The pool of test items is calibrated on the same scale as the estimates of test takers' ability. From this, a test taker's probability of correctly answering a question can be calculated even if that person did not answer the question. The IRT-estimated correct score is the sum of probabilities of correct answers for the pool of science questions. The IRT-estimated correct score can help correct for distortion of scores in cases of students correctly guessing on hard questions or omitting answers to questions. The results reported in this analysis are based on the IRT-estimated correct test score (NCES 1995).

Hands-on Science Scales

The NELS student and teacher questionnaires contain a number of items concerning science instruction in the classroom. For students, these items focus on the frequency of instructional practices, e.g., how often did you watch the teacher demonstrate an experiment. Student responses range from 1 to 5 with 1 being low and 5 high. For teachers, there are three types of items. Some items concern the frequency of certain practices that use a similarly coded 1 to 5 scale, e.g. how often the teacher demonstrates an experiment. Other items concern the percent of class time spent on certain practices, e.g. percent of time spent conducting lab periods. Teachers had the opportunity to answer on a 1 to 6 scale but none gave a response of 6, the highest response. Third, teachers were asked to comment on access, quality and quantity of equipment. While there are some similarities between the teacher and student items, they differ in many aspects.

A hands-on science scale was created separately for teacher reports and for student reports in each grade. Factor analysis was used to identify which of the teacher and student items concerning science instruction provided the main source of variation in the scale. Those items identified were summed and averaged, creating a scale with a value of 1-5 (1 being low).

The NELS items regarding the science instructional practices in the classroom did not stay constant over the three waves. In the base year, 8th grade, no items regarding science instruction were included in the student questionnaire. Therefore, we cannot construct a student scale for 8th grade preventing a direct comparison with the RAND 8th grade student scale. Instead, only a teacher scale can be constructed for the 8th grade data. Due to the addition of some student items for the 10th grade survey, we are able to create a 10th grade student scale. However, one item found key in the RAND analysis (how often do you conduct experiments) was not included in the student survey. Therefore, our NELS 10th grade student scale has an important difference from our RAND scale, giving us less confidence in the construct validity of the NELS scale and reducing the value of its comparison with the RAND student results. In the 12th grade student survey this key item (how often do you conduct experiments) was added and the student scale more closely resembles the RAND student scale.

Teacher items were modified as well over the three waves. For this reason, separate factor analyses were conducted for each wave, thus, the hands-on scales used in the analyses are not the same for each wave. However, the three teacher surveys contained a similar item asking how often students did experiments and all three teacher scales include this item. The scales used were chosen based on the level of their reliability[2].

Table 5-1 identifies the items used in making each hands-on scale and the reliability of each scale. In comparison, the RAND teacher scale also contains time spent on labs and frequency of use of materials which is somewhat similar to the frequency of experiments. The NELS teacher scale contains information on non-hands-on activities that may be related to the frequency of hands-on science (e.g. demonstrations or reports)

[2] As a check, we created other hands-on scales that reflected more uniformity between years. Results from these analyses will be discussed in the Results section.

which are not part of the RAND teacher scale. The RAND student scale focuses on the frequency of doing experiments. The NELS student scale varies a great deal between the 10th and 12 grade as the 10th grade lacks information on the frequency of experiments and focuses on possibly related non-hands-on activities. The 12th grade scale does include this information as well as the related activities.

The NELS student scale also differs from the teacher scale in its interpretation. The teacher scale is provided by science teachers having a sampled student in their class. Thus, all students who are linked to a teacher scale are currently (meaning in that year of the NELS) enrolled in science. However, the student survey asks for responses based on "the most recent science class". Since four years of science may not be required, a student's scale may be based on the level of hands-on science in a class from one or more years ago. This is most likely to occur for 12th graders. Due to this, the models for the NELS analysis assume that the student-reported level of hands-on science in a class taken this year versus a class taken one to several years ago will have the same relationship with 10th and 12th grade test scores. In the Sensitivity Analysis section (VII A) of this chapter this assumption is tested and found to hold.

Covariates

One of the strengths of the NELS data is the large number of variables collected, some of which have been found to be correlated with test scores in past research. In addition, these variables may also be correlated to the amount of hands-on science. If they are not included in the analysis, their relationship with test score may be confounded with the relationship found for hands-on science. If this occurs, we will not obtain an accurate measure of the relationship of hands-on science and test score. In Table 5-2 we identify the variables of this type included in the analysis to avoid this confounding problem. These variables address the individual, classroom and school level and are identified as such in the table. The first section of the table includes variables also used in the analysis of the RAND data and the second section contains the additional variables used in the analysis of all three waves of the NELS data.

For the covariates available in both the NELS and RAND data, there are some differences. The ability rank quintiles were constructed from teacher rankings in the RAND data. Since such rankings were not available in the NELS data, a composite of English and math grades was used to place students in rank quintiles (based on the grades of the entire NELS sample). The ability rank variable was constructed as quintiles to allow an analysis similar to that used with the RAND data and to check for non-linear effects (as were found in the RAND data)[3].

In addition, certain NELS covariates differ between grades. The Percent Minority variable was a classroom level variable in the 10th and 12th grade NELS data (and in the RAND data) but for 8th grade only a school level variable was available. Type of Science Classes Taken was not available for 8th grade, it was reported by students in 10th grade and it was based on transcripts for 12th grade. Amount of Science Taken was based on these same sources.

[3] The construction of a continuous Ability Rank variable and its effects on the results are discussed in the Results Section.

III. Descriptive Analysis

The descriptive analysis of the NELS data serves the dual purpose of determining whether it is similar to the RAND data and whether its several years of data are similar to one another. We find it similar to the RAND data in two respects. Student test scores significantly vary by groups (gender, race/ethnicity and ability rank) and there is variation in the student hands-on scale by ethnic group and ability rank but not by gender (except in 12th grade). The two data sets differ in other respects. In the NELS, there is greater variation in the teacher hands-on scale by group. Such variation enables us to estimate more precisely the role of hands-on science as reported by teachers.

As the NELS data is longitudinal, we can check for differences among the waves. The descriptive data presented here are weighted using student weights to address the two-stage cluster design and non-response in order to reproduce the population patterns. These weights include a weight representing the student's probability of selection and a student-level non-response weight based on the combination of school type, school region, ethnicity and gender. Cross-sectional weights were created for each grade and apply to all members of a wave regardless of their participation in any other wave (NCES 1994, 1995).

Table 5-3 provides the weighted univariate analysis of the data showing the means and standard deviations for the variables. As expected, we find that test scores rise over time as do their standard deviation: the mean of the test score increases with grade (from 18.5 to 21.8 to 24.06) as does the standard deviation (from 4.8 to 5.9 to 6.1). We would expect the rise in the mean as the tests contain many of the same questions from the wave before and student achievement should rise at different rates with additional years of education.

The mean teacher hands-on scale and its standard deviation are similar over the three waves. The student scale exhibits more variation between 10th and 12th grade which may in part be due to the inclusion of an important item on how often experiments are done in the 12th grade survey. The teacher hands-on science scale is similar throughout the three grades with a mean running from 2.69 to 2.75 on a scale of 1-5 with a standard deviation between .66 and .70 (Panel I). The student hands-on science scale is only available for 10th and 12th grades. The over time difference in the student scale (2.48 versus 2.75) is greater than in the case of the teacher scale (2.74 versus 2.75) (Panel I).

Concerning the other covariates, Table 5-3 shows stable values across grades except in two cases. The variables of Science Class Track and Class Achievement level change between years (Panel II). By 12th grade, a majority of students have been placed in high track and higher achievement by the teachers. The standard deviations for these variables remain fairly constant.

Table 5-4 provides the weighted bivariate analysis showing significant differences within certain groups for the hands-on scales and test score. The multiple choice test scores differ among groups in an expected pattern, similar to the one seen with the RAND data. Males score higher than females (Panel I). Asian and White students score higher than Black and Hispanic students (Panel I). Scores increase monotonically with SES quintile, ability rank quintile, and achievement level of the class (Panels II & III). Regarding the teacher hands-on scale we see no pattern of differences among the groups for Gender or Race (Panel I). There is a pattern that teachers of classes with students of higher SES, ability rank and classes of higher class achievement report more hands-on instruction (Panels II & III). This pattern hold true for all three grades. For SES, this pattern holds true for 8th and 10th grades (Panel II).

Regarding the student hands-on scale we do not see a similar pattern within the groups. Instead, we see opposite results between the two grades for several of the groups. For example Black reports the lowest scale in 10th but the highest in 12th grade (Panel I). In 10th grade, the highest ability rank students report a higher scale but in 12th grade the lowest report a higher scale (this is also true for class achievement) (Panels II & III). These anomalies suggest a problem with the reliability of the student hands-on science scale.

IV. The Models

As a first step, we will examine the NELS data using the RAND model (the model used to analyze the RAND data) to check if we get similar results. The RAND model contained covariates for gender, race/ethnicity, classroom percent minority and ability rank.

Our next step will be to make use of the additional covariates available in the NELS data and estimate a model using standardized multiple choice science tests for each of the three waves of data. The NELS model takes the following form:

1) $Y_{ij} = \alpha_0 + \alpha_1 H_{ij} + \alpha_2 ST_{ij} + \alpha_3 CL_j + \alpha_4 SCH_j + \varepsilon_{ij}$

where
Y_{ij} = the multiple choice test score for student i in class j.
H_{ij} = the level of hands-on science for student i in class j (the student or teacher hands-on science scale - as teacher reported data is at the classroom level it should be represented by H_j in the above equation).
ST_{ij} = the student characteristics (ability rank, gender, race/ethnicity, SES, hours of homework and science coursework) that may be related to test scores for student i in class j.
CL_j = the class level variables (percent of minority students, class track, and achievement level of class,) that may be related to test scores for class j.
SCH_j = the school characteristics (type, metropolitan status and regional location) that may be related to test scores for student i in class j.
α's = the parameters to be estimated (α_1 being the coefficient for the hands-on scale)
ε_{ij} = the disturbance term for student i in class j.

In addition, one extension and one modification were made to the NELS model. We extend the NELS model to address our finding of a significant effect for an interaction term between the ability rank and the student hands-on science scale that was seen in the RAND analysis. Similar interaction terms (using both the student and teacher hands-on scales) were introduced into the NELS model. These interaction terms were found to be non-significant so were not included in the NELS model.

We also modified the model to address the greater number of student reports versus teacher reports for the students in 10th and 12th grades. Only for a sub-sample in each wave were the students' science teachers surveyed. Therefore, we have more student reports and test scores than teacher reports which directly affects the three class level covariates provided by the teacher reports: percent minority of science class, science class track, and achievement level of students in the science class. The sample available for analysis when using these three covariates is much smaller than when not using them. The discussion in the section on the data (along with Tables A5-2 and A5-3) notes that the descriptive statistics of these two samples are similar. As a further check on the effects of this reduction in sample size, the NELS model was estimated both with and without the three class level variables for the 10th and 12th grade analysis of the student hands-on scale. This modification did not result in any important changes to our results.

In addition, for covariates that were missing over 100 observations, imputation was done by adding a dummy for missing observations while changing the missing values to 0. Imputation did not change the results regarding the relationship of the hands-on scales to test scores. For some of the covariates, it did increase the efficiency of the estimates.

V. Results

The results from the multivariate analysis are discussed below in three sections. First, using both the teacher and student scales the data is analyzed using the same model as used with RAND data. Second, the NELS model is estimated using the teacher hands-on scale. Third, the estimation of the NELS model is made using the student hands-on scale.

The RAND Model

We first analyze the NELS data using the same model we used with the RAND data to examine whether similar results can be obtained from both data sets. Compared to the model used for the NELS data this model contains fewer covariates (gender, race/ethnicity, ability rank, classroom percent minority and, where significant, the interaction of ability rank and the hands-on scale). Table 5-5 contains the results from the estimates of this model[4]. On the left side of the table are the results when using the teacher hands-on science scale and on the right side are the results when using the student hands-on scale. We show only the model with the best fit for each grade-scale combination. This is why the table contains the scale-ability rank interaction terms only for the 12th grade-teacher scale and the 8th grade-student scale. The R^2 for these models are similar and range from .23 - .35.

[4] The tables in this section show unweighted results as weights are not used in the multivariate analysis. In cases where weights are a function of the independent variables included in the model, unweighted OLS estimates have been found to be unbiased, consistent, and have smaller standard errors compared with the weighted OLS (Winship and Radbill 1994). We have included as independent variables in our model all the variables used to construct the NELS:88 student weights: school type, school region, student ethnicity and student gender. For this reason, we use unweighted analysis.

Beginning with our findings for the teacher hands-on scale ,we see very different results when using the NELS data. For the RAND data, the coefficient for the teacher scale was insignificant. For the NELS data it is positive and significant for all three waves of the NELS data and ranges from .75 to .92. In Chapter 4, we suspected that the small number of teachers and lack of variation in the hands-on science reports in the RAND data may have been responsible for the non-significant coefficient. The NELS data contains a much larger sample of teachers and greater variation in their hands-on science reports. The different findings for the RAND versus the NELS data may reflect this difference.

Using the RAND data, we found no significant interactions between the teacher hands-on scale and ability rank. From the NELS data for the 8th and 10th grade data, we did not find a significant coefficient for the interaction of teacher hands-on scale and ability rank either. For the 12th grade NELS data we find two significant negative coefficients for this interaction. The coefficients are large enough to offset the positive effect of the teacher scale for the higher ability ranks (Ranks 3 & 5 but not Rank 4). The results for 12th grade are similar to those found when using student reports in the RAND data in that for both cases the positive relationship of hands-on science and multiple choice test score was confined to students of lower ability rank.

Results for the coefficients on the other covariates are very similar for the RAND and the NELS data. We see negative coefficients for female (though only marginally significant for the RAND data), race/ethnic group (with consistent results for Black and Hispanic), and percent minority. We also find positive monotonic coefficients for ability rank.

Turning to the results for the student hands-on science scale (right side of Table 5-5), we do not have an 8th grade student scale from the NELS so the comparison is for 10th and 12th grade data only. Our findings differ from the RAND data results. The RAND data showed a positive significant coefficient for the student hands-on scale (for students of lower ability rank) as does the NELS 10th grade data but the NELS 12th grade data shows a negative significant coefficient. This multivariate result from the NELS mirrors the opposite patterns for grades 10 and 12 observed in the bivariate analysis. Futhermore, the interaction of the scale and ability rank was found significantly negative for the RAND data and large enough to offset the coefficient on the hands-on scale for students of high ability rank. This interaction was not found significant when using the NELS data and the coefficient on the hands-on scale applies to students of all ability ranks.

The coefficients on the covariates are very similar for the two data sets. Their signs and significance levels agree in all cases (negative for female, race/ethnicity, percent minority and positive for ability rank).

The NELS Model: The Relationship of the Teacher Scale and Test Score

The next step is to examine the results from the NELS model which contains additional covariates that may be associated with both the level of hands-on science and student test scores. We have greater confidence in these findings, than those using the previous model, as the addition of the covariates reduces possible confounding effects they may have had on the coefficient for hands-on scale when these covariates are not included. Table 5-6 shows the results for all three grades when estimating the model using the teacher hands-on scale. The imputed results shown on the right side of the table are very similar to the non-imputed ones on the left side (though some coefficients become significant when using the imputed data as expected) and we focus our discussion on the results from the imputed data. The R^2 for these models run from .34 - .43, an improvement upon that seen in the RAND model. Since the interaction between the hands-on scale and ability rank was not found significant, for the sake of parsimony the final NELS model does not contain it.

In regards to the relationship of the teacher hands-on scale and test scores, the results for the 8th and 10th grades are very similar. In both cases, the coefficient for the hands-on scale is positive and significant with a similar magnitude of .43 and .32 respectively (Panel I). The additional covariates have reduced the magnitude of this coefficient (from what is seen in Table 5-5) but have not changed its sign nor significance. These coefficients can be converted into standard deviation units of test score which are .09 and .05 respectively.

The 12th grade results differ from those of the other two grades and the results when estimating the RAND model. The coefficient of the hands-on scale is insignificant and changed from its previous positive significant value by the addition of the covariates.

The results for the covariates are similar among all three grades. Negative significant coefficients are seen with Female, Black, Hispanic, and Percent Minority. Positive significant coefficients are seen with higher ability rank[5], SES, Class Track, Achievement Level and non-religious private school.

[5] The Ability Rank variable was constructed as quintile dummy variables. To check if this construction affected its results, the model was also re-estimated using a continuous ability rank term and an interaction term composed of the hands-on scale and continuous ability rank. For all grades and for both the student and teacher scale:

1. The continuous ability rank variable had a positive significant coefficient
2. The coefficient on hands-on scale was not affected
3. The interaction term was not significant.

In sum, our results are similar when using a quintile ability rank or a continuous ability rank variable.

The NELS Model: The Relationship of the Student Scale and Test Score

Next, we estimated the model using the student hands-on scale for the two waves (10th and 12th grades). As a modification, the model was estimated without the teacher-reported classroom level variables to maintain a larger sample. Table 5-7 shows the results for the 10th grade of the model and its modification when using non-imputed and imputed data. Table 5-8 shows the same for the 12th grade. The R^2 for these models run from .36 - .46, an improvement on the RAND model. For both grades, there is little difference in the results between the model and its modification and little difference between the non-imputed versus imputed data within grades. We focus our discussion below on the unmodified model using the imputed data.

There are major differences between the results for the two grades. For the 10th grade, the coefficient for the hands-on scale is not significant (Table 5-7, Panel I). This finding is in contrast both to the finding from the RAND data and from estimating the RAND model using the NELS data where the results showed a positive significant coefficient. The inclusion of the additional covariates have changed the original results.

For the 12th grade, the coefficient is negative and significant (Table 5-8, Panel I). This finding is similar to that when using the NELS data with the RAND model but very different from our findings for the RAND data. This result may be due to the relatively unreliable data on student hands-on science of the NELS.

The results for the covariates are similar for the two grades and with the results from the models using the teacher hands-on scale. In addition, we see significant positive coefficients for geographic regions when compared to the South.

VI. Discussion

Our findings from the cross-sectional analysis of the NELS data are more of a compliment to the analysis of the RAND data rather than a confirmation. On the one hand, we find evidence for a positive association between the teacher hands-on scale and test scores that we did not find in the RAND data. On the other hand, we did not find evidence to support such an association between the student hands-on scale and test score using the NELS data.

In addition, we did not find evidence supporting the importance of the interaction of hands-on scales and ability rank as was found with the RAND data when using the student scale. We did have a similar finding when estimating the RAND model when using the 12th grade data with the teacher scale but this disappeared in the NELS model.

One concern with our finding of no relationship between the teacher scale and test scores in the RAND data was the small number of teachers involved and the lack of variation in their reports on the amount of hands-on science in their classes. Using the NELS data with its larger teacher sample and the significant variation in the scale for different groups (e.g. ability rank, class achievement and SES) addresses this concern. Our findings of a positive relationship between the teacher scale and test scores for the 8th and 10th grades give us greater confidence that the lack of variation in the teacher scale in the RAND data causes the failure to find such a relationship.

Our failure to find a positive significant relationship when using the student hands-on scale for 10th grade and our finding of a negative relationship for 12th grade when using the NELS data seemingly contradicts our findings from the RAND data. At the same time, this finding is open to several interpretations. The NELS data does not contain a student scale for 8th grade so no direct grade level comparison could be made with the RAND analysis. The 10th grade scale suffers from a lack of appropriate items. Thus we do not have the tools to make a full comparison between the two data sets as we were able to do for the teacher hands-on scale. Additionally, hands-on science may not be as effective in 12th grade as students are more capable of abstract thinking (and this may explain our finding no relationship for 12th grade using the teacher scale as well).

These results reveal the complimentary nature of the two data sets. The RAND data is stronger regarding the student scale and the NELS data is stronger regarding the teacher scale. Our results would be expected to reflect the strengths and weaknesses of the data used.

VII. Sensitivity Analysis

Further analyses were carried out to investigate the sensitivity of our results using the NELS data. Two topics were investigated: 1) the measurement of the teacher scale and 2) measurement of the timing and order of science course taking

.

The Measurement of the Teacher Scale

The hands-on scale is intended to combine different dimensions of a broader construct of hands-on science. For this reason the scale is made from several items measuring different topics, e.g. time spent in labs; frequency of student experiments, teacher demonstrations or student reports on experiments; and amount and condition of scientific equipment. Only the 8th grade teacher scale is made up of all these different items. The 10th and 12 grade teacher scales combine time spent in labs with frequency of student experiments and reports with the 10th grade scale also including frequency of teacher demonstrations. The use of different teacher items for each grade's scale was based on the results of our factor analysis plus the choice of the scale with the highest level of reliability for each grade.

In order to check the robustness of the results using the original teacher scale, we estimated the models using several different items alone or in combination. These included:

1. Time spent conducting lab periods
2. Frequency of student experiments
3. Time conducting lab periods and frequency of student experiments
4. Frequency of student experiments and frequency of student reports on experiments
5. Amount of scientific equipment and condition of scientific equipment

The purpose of scales 1,2,3 and 5 is to use a scale composed of the same items in all three grades. Scale 4 uses the same items available in grades 10 and 12. Scale 5 also addresses hands-on science in a different manner than the other four scales strictly through the consideration of scientific equipment. The NELS model was estimated with each of these five scales. In Table 5-9 the resulting coefficients on the scales are compared to the coefficients on the original scales. Our original results are robust when compared to the alternative hands-on scales. For 8th grade, the coefficients on the alternative scales are all positive and significant though their magnitude is somewhat smaller. For 10th, the coefficients on the alternatives are again positive and significant and their magnitudes are almost all similar to the coefficient on the original scale. In 12th grade, the coefficients swing from positive to negative but none are significant reflecting the lack of significance of the coefficient on the original scale.

If we had found that the modified scales for the 8th and 10th grades were not positive and significant, we might have ascribed the lack of significance in the 12th grade original scale as due to a mis-specified scale, as it was made up of fewer items than the scales in the other two years. Instead, we found that the results were similar for the alternative scales for 8th and 10th grade. Therefore, our finding of no relationship between hands-on science and test score in 12th grade might be attributed to a difference in hands-on science's relationship to test score for older, more advanced students.

Timing and Order of Science Course Taking

Student course taking in science has been linked to student standardized test scores (Mullis & Jenkis 1988, Jones et al. 1992). Much of the explanation for this relationship can be classified as opportunity to learn. Course taking may also reflect student ability and the inclusion of an ability variable in the model reduces this aspect, leaving topic coverage as the most likely explanation for its relationship with test score. If students are not exposed in school to materials covered on a test they are less likely to have learned them and so are less likely to answer correctly. The NELS science test is constructed in a manner that increases the importance of opportunity to learn. To avoid ceiling effects, the version of the science test was changed for each grade. Along with the addition of more difficult items, items based on material expected to be covered only in the later grades were added to each successive version. So if students do not take the courses in which this material is expected to be covered, they are less likely to be able to correctly answer the added items.

If course taking is also related to the level of hands-on science but is not considered, we may confound the effect of course taking with that of hands-on science. Specifically, if we do not address course taking in our model, we will have omitted variable bias as the effect of course taking will be taken up by the error term. But as course taking is related to hands-on science, the error term in an equation excluding course taking will correlate with hands-on science. Our assumption of zero covariance between hands-on science and the error term will not hold and the coefficient for hands-on science may be biased.

To avoid this problem, our NELS model includes as covariates both the type of science courses taken and the number of courses taken. The inclusion of these variables did not change the sign or significance of our results regarding the hands-on scale although it did slightly reduce their magnitude. The coefficients on these covariates are significant and so they remain as a part of the NELS model. Table 5-10 reviews the coefficients (from tables 5-6 to 5-8). The coefficients on the individual subjects are in comparison to taking "other science courses". The results may reflect that the NELS test is made increasingly difficult by the addition of new material which in high school includes more physics and less biology. We see a decline in biology's positive relationship to test score from 10th to 12th grade, a sign change in physics and a stable relationship for chemistry. An unexpected finding is that the number of total courses has a negative relationship in 10th grade and but positive relationship in 12th grade. Possibly these results are due to the use of student reports in 10th grade and a transcript review in 12th grade.

There are two additional explorations that can be made into course taking and the possibility of confounding effects: 1) current enrollment in a science course, and 2) the pattern of science course taking during secondary school.

A. Current Enrollment in Science

The issue of current enrollment concerns the student hands-on scale. The items that make up the scale are asked in the following manner: "In your current or most recent science class, how often did you....?". When using the student hands-on scale, our sample may include students currently in science class and those who may not have taken a science class for 1 term to 3 years (the latter can be true for 12th graders who have not taken a science class since 9th grade). There may be a relationship between the recentness of taking a science course and a student's test score and a relationship between this recentness and the level of hands-on science. If so, we again face the problem of omitted variable bias. In addition, we may face a reliability problem as students who are not in a science course may be less likely to correctly remember the level of hands-on science in their last science class.

The issue of current course taking is not a concern when using the teacher hands-on scale. The teacher scale is provided by a sample of science teachers who have a NELS student in their science class when surveyed. Only students currently enrolled in a science class are therefore considered when estimating the model using the teacher scale.

For the student hands-on scale, this issue is addressed in two ways. First, when estimating the NELS model using the student hands-on scale, a dummy variable on student enrollment in a science course is included. This variable is based on student response to an item regarding the level of science homework. One of the responses is "not taking science". This variable was not found to be significant (although it was marginally significant when using the 12th grade data). As an additional check, we estimated the NELS model with an additional interaction term composed of the student scale and the "not taking science" dummy variable. The coefficient on the interaction term would show whether not being in a science class this term led to a different relationship between the hands-on scale and test score. For example, a significant negative coefficient would reflect a weaker relationship for students not presently taking a science course. For both the 10th and the 12th grade, neither the interaction terms nor the "not taking science" dummy was significant. Therefore, we found no evidence of a difference in the relationship of hands-on science to test score due to current enrollment in a science class.

B. The Pattern of Science Course Taking

The NELS model controls for each type of course taken. A further step would be to examine whether the pattern of types of science courses an individual student takes affects the relationship of hands-on science to test scores. Possibly, the pattern may be correlated with both test scores and the level of hands-on science, creating the potential for omitted variable bias.

The following discussion looks first at 10th grade course taking patterns then 12th grade patterns. We examine course taking in each grade using four steps. First we identify the top five course taking patterns. Second, we examine the level of hands-on science for each pattern. Third we examine the coefficient for hands-on science when the NELS model is estimated using only the students who fit within one pattern. Fourth, we modify the NELS model to include all the patterns and their interaction with the hands-on scale to test if they affect the relationship of the hands-on scale to test score.

1) 10th Grade

Tenth grade students were asked to report on their science courses taken in 9th and 10th grade (be they semester or year long courses). They were not asked to identify the order of the courses taken but because biology is often required in the 9th grade and the majority of students reported taking biology, we can assume that biology was taken first for the majority of students reporting it. Table 5-11 lists the top 5 patterns of course taking as self-reported by 10th grade students. One reason that the Biology, Other Science pattern is the largest is due to Other Science being a catch-all category of courses rather than one specific course.

Our second step is to examine whether the level of hands-on science for each pattern varies from that of the whole sample. Table 5-12 provides the mean hands-on scale (both teacher and student) for the whole sample and each of the top 5 course patterns.

A superscript identifies a significant difference from the mean of another pattern. For example, under the teacher scale we see that all the patterns have a 2 superscript (except for B,C) which shows that the mean of B,C is significantly greater than that of the other patterns. This means that the current teachers of B,C students have reported a higher level of hands-on science in their classrooms. Looking at the student scale, we see that students with a B,C pattern report significantly more hands-on science in their current or last class than students of other patterns while students with a B,O pattern report significantly less hands-on science.

Our third step is to estimate the model separately for students within each course pattern. Table 5-13 shows the coefficient on the hands-on scales for the NELS model using the whole sample and for each course pattern. For the teacher scale, we see two patterns (marked in bold) that appear different from our original finding of a positive significant coefficient. For the pattern B (biology only) we find a negative though not significant coefficient. For the pattern B,E (biology and earth science) we find a positive significant coefficient but its magnitude is more than double that of our original finding.

Interestingly, for the pattern B,C which showed a higher mean value of hands-on science than the other patterns (see the table above) we see a coefficient similar to that when using the whole sample.

For the student scale, we see two patterns (marked in bold) that appear different from our original finding of a negative non-significant coefficient on the hands-on scale. For B,E (biology and earth science) we see a positive but not significant coefficient. For the pattern B,O (biology, other science) which had a significantly lower mean value of hands-on science (see table above) we see a negative significant coefficient. Again, for the pattern B,C which showed a higher mean value of hands-on science than the other patterns (see the table above) we see a coefficient similar to that when using the whole sample.

We want to investigate whether or not differences in the coefficient for the hands-on scale between the whole sample and the patterns are significant. To test this we estimate an extension of the NELS model for the whole sample which includes all the pattern dummy variables plus an interaction term composed of the pattern in question and the hands-on scale. For the teacher scale, there are two patterns in question (B and B,E) and for the student scale two patterns as well (B,E and B,O). The model is estimated four times using an interaction term for each of these patterns one at a time. If the interaction term is significant, then we will conclude that there is a significant difference in the relationship of hands-on science to test scores for that specific pattern versus the whole sample. Tables 5-14 and 5-15 show the results from estimating this model.

First, we examine the teacher hands-on scale and its interaction with two patterns, B and B,E, which appear to differ from the whole sample. Table 5-14 shows the coefficient on the hands-on scale for the NELS model using the whole sample and for the extended models which include the pattern variable and the interaction of the pattern and the hands-on scale for the 10th graders.

The coefficient on the teacher hands-on scale is similar for all three models. While the coefficient for the interaction term with B is not significant, the coefficient on the interaction term for the pattern B,E is significant and positive (.79). Students taking the course pattern (biology and earth science) have a stronger relationship between the teacher hands-on scale and their test scores.

Second, we test the student hands-on scale and its interaction with two patterns, B,E and B,O, which appear to differ from the whole sample. Table 5-15 shows the coefficient on the hands-on scale for the NELS model and for the extended models which include the pattern variable and the interaction of the pattern and the hands-on scale for the 10th graders.

For the model containing the interaction of B,E and the student scale, the coefficient on the student scale remains not significant and similar to that of the original model. The coefficient on the pattern variable is positive and significant and the coefficient on the interaction term is not significant when the class level covariates are included but is significant when they are left out

For the model containing the interaction of B,O and the student scale, the coefficient on the student scale becomes positive and marginally significant. The coefficient on the pattern variable is positive and significant. The coefficient on the interaction term is negative and significant. We combine the coefficient on the hands-on scale and the coefficient on the interaction term: .11 + (-.35) = -.24. Using a test of equivalence of coefficients (H_o: Coefficient on Main Scale + Coefficient on Interaction Term = 0), we find that we can reject the hypothesis that the resulting final coefficient equals 0 at the 5% significance level. Students with the B,O pattern of course taking have an overall negative relationship of hands-on science and test score.

To sum up the results for the 10th grade, we find that course pattern does not affect our result of a positive relationship between the teacher hands-on scale and test score but that the relationship is increased for course pattern biology-earth science (B,E). For the student hands-on scale, we find a negative significant relationship with test score for the specific pattern of biology-other science (B,O).

2) 12th Grade

The NCES gathered 12th grade transcripts from the sampled students and from this constructed variables on the courses taken and number of Carnegie units in each course. These variables have the advantage of not relying on student memory as the 10th grade variables do. Like the 10th grade variables, they do not show the order in which the courses were taken. Table 5-16 lists the top 5 patterns of course taking as gathered through a review of 12th grade student transcripts. The Biology, Other Science pattern remains the largest pattern though it now makes up one-fifth of the sample rather than one-third as it did in 10th grade. The number of students with B,O status is furthered reduced by the larger percentage of missing data for the hands-on scales for students in this pattern (see n's in Table 5-17).

We examine whether the level of hands-on science for each pattern varies from that of the whole sample. Table 5-17 provides the mean hands-on scale (both teacher and student) for the whole sample and each of the top 5 course patterns.

There are no significant differences regarding the mean of the teacher scale among the five patterns. Looking at the student scale, we see that students with a B,C,O pattern and, to a lessor extent, the B,C pattern report significantly more hands-on science in their current or last class than students with B,C,P and B,C,O,P patterns. The differences in the mean student hands-on scale among the patterns and between the patterns and whole sample is much less than what was seen with the 10th grade data.. We do not see any significant differences with the B,O pattern as we saw in 10th grade.

Next, we estimate the model separately for students within each course pattern. Table 5-18 shows the coefficient on the hands-on scales for the NELS model using the whole sample and for each course pattern. For the teacher scale, we see two patterns (marked in bold) that appear different from our original finding of a positive non-significant coefficient. For the pattern B,C (biology, chemistry) we find a negative though not significant coefficient. For the pattern B,C,O, P (biology, chemistry, other and physics) we find a positive marginally significant coefficient but its magnitude almost six times that of our original finding.

For the student scale, our results for the individual patterns are very similar to those of the original finding. In all cases, we see a negative significant coefficient.

To test whether or not differences in the coefficient on the hands-on scale between the whole sample and the course patterns are significant, we estimate an extension of the NELS model as we did for the 10th grade data.

For students with the patterns B,C and B,C,P,O the results appear to differ from that of the whole sample. Table 5-19 shows the coefficient on the teacher hands-on scale for the NELS model using the whole sample and for the extended models which include the pattern variable and the interaction of the pattern and the hands-on scale.

The coefficient on the teacher hands-on scale is not significant for any of the three models. The coefficient on either of the interaction terms is not significant so we conclude that course taking pattern does not change our original results for 12th graders.

Because the results for the different patterns were similar to those of the whole sample when using the student scale, we do not expect to find a significant interaction effect that differs from the whole sample. Our tests (not presented here) show that for four of the patterns we found no significant interaction effect and no change in the sign and significance of the coefficient on the student hands-on scale. For the pattern B,O, we continued to find a negative significant coefficient on the student scale but also a positive significant coefficient on the interaction term. The combination of the coefficients of the student scale (-.57) and interaction term [-.57 + .30 = (-.27)] remained negative and significantly different from zero.

In sum, the analysis of the 12th grade data shows that course pattern does not affect our results regarding the hands-on scale when using either teacher or student scales. As in 10th grade, taking courses in the pattern B, O seems to reduce but not dispel the negative relationship found between the student scale and test score.

Summary of Sensitivity Analysis

The robustness of the results of our sensitivity analysis increases our confidence in the findings from the NELS model. Different compositions of the scale did not change our findings of a significant positive relationship between the teacher scale and test score.

Modifying the NELS model to include greater detail on the timing and order of science course taking also provided findings similar to the NELS model. The NELS model's inclusion of course type and number did not change the sign nor significance of the coefficient on the hands-on scale (though the magnitude was reduced). The inclusion of current enrollment in science was found not to be significant nor did it change the coefficient of the hands-on scale.

On the whole, inclusion of the pattern of course taking also did not change the findings of the NELS model. For the teacher scale, only with one of the five top 10th grade course-taking patterns was there a significant change in the results (the coefficient on the interaction term was the same sign as the coefficient on the teacher scale) which did not change the conclusions drawn from the original results. For 12th grade, course taking patterns had no significant impacts on the original result for the hands-on scale. For the student scale, the introduction of course taking patterns into the model led to no significant change in the results with the one exception for 10th grade. Overall then, our sensitivity tests hold with the results of the NELS model and provide us with greater confidence that we can base conclusions upon them.

Table 5-1: Composition and Reliability of Hands-on Science Scales Used in NELS Analysis

Grade	Type of Scale	Items	Alpha Reliability
8th	Teacher	1. Time spent conducting lab periods 2. How often teacher demonstrates an experiment 3. How often student experiments are done 4. Access to lab in classroom 5. Amount of scientific equipment 6. Condition of scientific equipment	.80
10th	Teacher	1. Percent time spent conducting lab periods 2. How often teacher demonstrates an experiment 3. How often experiments are done 4. How often reports on experiments are done	.74
10th	Student	1. How often watch teacher demonstrate an experiment 2. How often write reports of lab work 3. How often use books to show how experiments work	.62
12th	Teacher	1. Percent time spent conducting lab periods 2. How often have students do experiments 3. How often require reports on experiments	.74
12th	Student	1. How often watch teacher demonstrate an experiment 2. How often do experiment alone or in group 3. How often use books to show how to do an experiment 4. How often write reports on experiments and observations	.77

Table 5-2: Covariates Used in the Analysis of the NELS Data

Analysis	Level	Variable	Values
NELS & RAND			
	Student	Gender	Female & Male
		Race/Ethnicity	Asian, Black, Hispanic, White
		Ability Rank	Quintiles (1 is low)
	Classroom	Percent Minority	RAND: 0 – 100% NELS: Scale of 0-6 with 0 = 0% and 6 = 91-100%
NELS Only			
	Student	Family SES	Standardized scale with a mean of 0 and std. dev. of 1
		Hours of Science Homework	Scale of 0-3 (0 meaning 0 hours and 3 meaning four or more hours a week)
		Types of Science Classes Taken	Dummies (0,1) for biology, chemistry, earth science, physics and other science (for 10th and 12th grades only)
		Amount of Science Taken	10th grade: number of different science courses taken in 9th & 10th grades 12th grade: Carnegie units for science taken in high school
	Classroom	Science Class Track	High versus Low
		Achievement Level of Students in Class vs. Average Student	Low, Average, High, Differential
	School	School Type	Public, Catholic, Private-other religion, Private-non-religious
		Metropolitan Status	Urban, Suburban, Rural
		Regional Location	Northeast, North Central, South, West

Table 5-3: Descriptive Statistics of Variables Used In NELS Analysis (weighted means and std. deviations)

Panel I

Variable	8th Grade	10th Grade	12th Grade
Individual Level Variables			
Science Test Score	18.52 (4.84)	21.81 (5.91)	24.06 (6.10)
Teacher Reported Hands-on Science Scale (1-5 with 1 low)	2.69 (0.66)	2.74 (.70)	2.75 (0.66)
Student Reported Hands on Science Scale (1-5 with 1 low)	NA	2.48 (.90)	2.75 (0.79)
Female %	0.50 (0.50)	0.51 (0.50)	0.50 (0.50)
Race/Ethnicity (White as reference) %			
Asian	0.03 (0.17)	0.04 (0.19)	0.05 (0.21)
Black	0.12 (0.32)	0.12 (0.32)	0.11 (0.31)
Hispanic	0.09 (0.29)	0.09 (0.29)	0.10 (0.30)
White	0.76 (0.43)	0.75 (0.43)	0.75 (0.44)
Ability Rank Quintiles (based on English & math grades) %			
Ability Rank 1 (low)	0.20 (0.40)	0.19 (0.39)	0.19 (0.40)
Ability Rank 2	0.18 (0.38)	0.18 (0.39)	0.20 (0.40)
Ability Rank 3	0.21 (0.41)	0.22 (0.41)	0.21 (0.41)
Ability Rank 4	0.23 (0.42)	0.22 (0.42)	0.20 (0.40)
Ability Rank 5	0.18 (0.38)	0.18 (0.39)	0.19 (0.40)
Family SES (standardized to a mean of 0)	-0.08 (0.76)	0.04 (0.76)	0.09 (0.77)
Hours of Science Homework (scale of 0-3 with 0 =no hours and 3 = 4 hours or more	1.05 (.64)	1.25 (.89)	1.65 (.97)

table 5-3 continued

Table 5-3

Panel II

Variable	**8th Grade**	**10th Grade**	**12th Grade**
Science Courses Taken (%)	NA		
(self reported in 10th grade and based on transcripts for 12th)			
Biology		.87	.95
		(.31)	(.21)
Chemistry		.18	.64
		(.39)	(.48)
Earth Science		.28	.22
		(.45)	(.41)
Physics		.04	.30
		(.20)	(.46)
Other Science		.64	.60
		(.48)	(.49)
Carnegie Units in Science	NA	NA	3.04
(scale of 0-11)			(1.05)
Classroom Level Variables			
School/Class Minority Composition	2.81	2.64	2.59
(scale of 0-7 with 0 low)	(2.07)	(2.21)	(2.11)
Science Class Track %			
High Track	0.25	0.58	.73
	(0.43)	(0.49)	(.44)
Other Track	0.75	0.42	0.27
	(0.43)	(0.49)	(0.44)
Achievement Level of this Class			
Versus the Average Student in the Grade %			
Differential Achievement	0.22	0.10	0.09
	(0.41)	(0.31)	(0.29)
Low Achievement	0.15	0.15	0.07
	(0.36)	(0.36)	(0.25)
Average Achievement	0.39	0.44	0.25
	(0.49)	(0.50)	(0.44)
High Achievement	0.24	0.31	0.58
	(0.43)	(0.46)	(0.49)

table 5-3 continued

Table 5-3

Panel III

Variable	**8th Grade**	**10th Grade**	**12th Grade**
School Level Variables			
School Control %			
Public	0.88	0.90	0.90
	(0.33)	(0.29)	(0.30)
Catholic	0.08	0.06	0.06
	(0.26)	(0.24)	(0.24)
Private - other religious	0.03	0.02	0.02
	(0.18)	(0.14)	(0.14)
Private - non-religious	0.01	0.01	0.02
	(0.12)	(0.11)	(0.13)
Metropolitan Status %			
Urban	0.23	0.27	0.26
	(0.42)	(0.45)	(0.44)
Suburban	0.43	0.42	0.41
	(0.50)	(0.49)	(0.49)
Rural	0.33	0.31	0.33
	(0.47)	(0.46)	(0.47)
Geographic Location %			
Northeast	0.18	0.19	0.20
	(0.38)	(0.39)	(0.40)
North central	0.27	0.27	0.27
	(0.44)	(0.44)	(0.44)
South	0.36	0.36	0.35
	(0.48)	(0.48)	(0.48)
West	0.19	0.18	0.19
	(0.39)	(0.39)	(0.39)
N			
Student reported variables	10,100	15,200	10,400
Teacher reported variables		5,400	3,000

Table 5-4: Scales and Test Scores By Selected Demographic and Student-Class Features (significant differences between groups shown in superscript)

Panel I

	8th Grade		10th Grade			12th Grade		
Independent Variable	**Teacher Scale**	**Test Score**	**Teacher Scale**	**Student Scale**	**Test Score**	**Teacher Scale**	**Student Scale**	**Test Score**
Total	2.69 (.66)	18.52 (4.84)	2.74 (.70)	2.48 (.90)	21.81 (5.91)	2.75 (.66)	2.75 (.79)	24.06 (6.10)
Gender								
Female	2.72 (.66)	18.27^{M} (4.62)	2.75 (.71)	2.49 (.88)	20.96^{M} (5.54)	2.76 (.64)	2.73^{M} (.80)	23.06^{M} (5.89)
Male	2.71 (.66)	19.21^{F} (5.24)	2.73 (.71)	2.48 (.91)	22.71^{F} (6.14)	2.74 (.69)	2.77^{F} (.78)	25.06^{F} (6.15)
Race/Ethnicity								
Asian	$2.93^{B,H,W}$ (.60)	$19.98^{B,H}$ (5.36)	2.81^{H} (.68)	$2.67^{B,W}$ (.86)	$22.67^{B,H}$ (6.15)	2.71 (.63)	2.77^{B} (.75)	$24.44^{B,H,W}$ (6.22)
Black	$2.56^{A,H,W}$ (.69)	$15.66^{A,H,W}$ (.3.79)	2.69^{H} (.77)	$2.50^{A,H}$ (.98)	$17.65^{A,H,W}$ (4.74)	2.69 (.69)	$2.88^{B,W}$ (.86)	$18.97^{A,H,W}$ (5.14)
Hispanic	$2.65^{A,B,W}$ (.68)	$16.55^{A,B,W}$ (4.19)	$3.03^{A,B,W}$ (.75)	$2.61^{B,W}$ (.95)	$18.94^{A,H,W}$ (5.00)	2.75 (.68)	2.86^{W} (.82)	$20.90^{A,B,W}$ (5.70)
White	$2.74^{A,B,H}$ (.65)	$19.51^{B,H}$ (4.90)	2.71^{H} (.68)	$2.45^{A,H}$ (.87)	$22.79^{B,H}$ (5.76)	2.76 (.66)	$2.71^{B,H}$ (.78)	$25.23^{A,B,H}$ (5.72)

table 5-4 continued

Note: Table reports means with standard deviations in parentheses.

The superscripts above the mean note which groups significantly differ ($p = .05$) from the group in question. For example, the mean for female test score in 8th grade reads 18.27^{M}. This superscript notes that the mean significantly differs from the mean male test score.

Table 5-4

Panel II

	8th Grade			10th Grade			12th Grade	
Independent Variable	**Teacher Scale**	**Test Score**	**Teacher Scale**	**Student Scale**	**Test Score**	**Teacher Scale**	**Student Scale**	**Test Score**
Family SES								
Quartile 1 (low)	$2.56^{2,3,4}$ (.71)	$15.93^{2,3,4}$ (3.83)	$2.66^{3,4}$ (.78)	$2.42^{3,4}$ (.93)	$18.75^{2,3,4}$ (5.02)	2.69^{4} (.67)	2.79 (.86)	$20.32^{2,3,4}$ (5.44)
Quartile 2	$2.63^{1,3,4}$ (.66)	$17.74^{1,3,4}$ (4.50)	2.69^{4} (.73)	$2.37^{3,4}$ (.92)	$20.72^{1,3,4}$ (5.62)	2.70^{4} (.68)	2.76 (.79)	$22.91^{1,3,4}$ (5.74)
Quartile 3	$2.71^{1,2,4}$ (.64)	$18.88^{1,2,4}$ (4.65)	$2.74^{1,4}$ (.66)	$2.51^{1,2,4}$ (.89)	$22.23^{1,2,4}$ (5.60)	2.76 (.68)	2.73 (.78)	$24.29^{1,2,4}$ (5.88)
Quartile 4 (high)	$2.90^{1,2,3}$ (.60)	$21.37^{1,2,3}$ (4.89)	$2.83^{1,2,3}$ (.63)	$2.58^{1,2,3}$ (.83)	$24.81^{1,2,3}$ (5.51)	$2.80^{1,2}$ (.63)	2.73 (.76)	$26.84^{1,2,3}$ (5.51)
Ability Rank Quintiles								
Ability Rank 1 (low)	2.67^{3-5} (.67)	$16.42^{2,3,4,5}$ (4.22)	2.61^{2-5} (.73)	2.36^{2-5} (.93)	19.24^{2-5} (5.16)	$2.61^{3,4,5}$ (.76)	2.79^{4} (.85)	20.38^{2-5} (5.510)
Ability Rank 2	2.66^{3-5} (.69)	$17.26^{1,3,4,5}$ (4.30)	$2.71^{1,5}$ (.68)	$2.44^{1,4,5}$ (.92)	$20.39^{1,3-5}$ (5.46)	2.71 (.68)	$2.83^{3,4,5}$ (.83)	$21.94^{1,3-5}$ (5.60)
Ability Rank 3	$2.75^{1,2}$ (.65)	$18.64^{1,2,4,5}$ (4.66)	2.74^{1} (.70)	$2.46^{1,4,5}$ (.90)	$21.24^{1,2,4,5}$ (5.31)	2.74^{1} (.69)	$2.76^{2,5}$ (.78)	$23.77^{1,2,4,5}$ (5.59)
Ability Rank 4	$2.75^{1,2}$ (.66)	$19.51^{1,2,3,5}$ (4.83)	2.79^{1} (.69)	2.56^{1-3} (.87)	$22.79^{1-3,5}$ (5.72)	2.82^{1} (.60)	$2.72^{1,2,5}$ (.76)	$25.69^{21-3,5}$ (5.45)
Ability Rank 5 (high)	$2.76^{1,2}$ (.63)	$21.71^{1,2,3,4}$ (5.03)	$2.82^{1,2}$ (.69)	2.58^{1-3} (.83)	25.51^{1-4} (5.88)	2.80^{1} (.62)	2.65^{1-4} (.73)	28.60^{1-4} (4.79)

table 5-4 continued

Table 5-4

Panel III

	8th Grade			10th Grade			12th Grade	
Independent Variable Achievement Level of Students in Science Class	**Teacher Scale**	**Test Score**	**Teacher Scale**	**Student Scale**	**Test Score**	**Teacher Scale**	**Student Scale**	**Test Score**
Differential Achievement within Class	2.69^{H} (.66)	$18.34^{H,L}$ (4.77)	$2.68^{H,L}$ (.73)	$2.36^{A,H}$ (.92)	$20.92^{H,L}$ (5.78)	$2.59^{A,H}$ (.78)	$2.53^{H,L}$ (.79)	$23.46^{H,L}$ (5.91)
Low Achievement	2.6^{H} (.66)	$16.30^{A,D,H}$ (4.28)	$2.56^{A,D,H}$ (.74)	2.42 (1.00)	$18.20^{A,D,H}$ (5.16)	$2.47^{A,H}$ (.73)	$2.84^{A,D,H}$ (.98)	$21.37^{A,D,H}$ (4.91)
Average Achievement	2.71^{H} (.67)	$18.34^{H,L}$ (4.66)	$2.76^{H,L}$ (.69)	2.48^{D} (.89)	$21.44^{H,L}$ (5.50)	$2.73^{D,H,L}$ (.67)	2.66^{L} (.80)	$23.23^{H,L}$ (5.75)
High Achievement	$2.79^{A,D,L}$ (.65)	$21.20^{A,D,L}$ (4.98)	$2.83^{A,D,L}$ (.67)	2.50^{D} (.85)	$24.61^{A,D,L}$ (5.54)	$2.82^{A,D,L}$ (.62)	$2.68^{D,L}$ (.74)	$27.29^{A,D,L}$ (5.35)

Table 5-5: Results From RAND Model

	Teacher Scale				Student Scale		
Variable	RAND 8th Grade	NELS 8th Grade	NELS 10th Grade	NELS 12th Grade	RAND 8th Grade	NELS 10th Grade	NELS 12th Grade
Teacher Hands-on Scale	-.43	.92**	.91**	.75*			
	(.37)	(.07)	(.10)	(.30)			
Student Hands-on Scale					1.28**	.27**	-.42**
					(.31)	(.05)	(.06)
Female	-.74^{m}	-1.25**	-2.34**	-2.41**	-.89*	-2.22**	-2.87**
	(.42)	(.09)	(.14)	(.18)	(.41)	(.18)	(.10)
Race/ethnicity (White as reference)							
Asian	-2.44**	-.30	.68*	-.22	-2.46**	-.37*	-1.15**
	(.56)	(.19)	(.30)	(.34)	(.57)	(.17)	(.19)
Black	-5.19**	-2.75**	-2.85**	2.93**	-5.00**	-4.17**	-3.91**
	(.62)	(.16)	(.28)	(.38)	(.62)	(.15)	(.18)
Hispanic	-3.05**	-1.99**	-2.24**	-2.19**	-2.66**	-3.26**	-2.95**
	(.59)	(.15)	(.26)	(.36)	(.55)	(.14)	(.16)
Missing Race/Ethnicity	NA	-2.69**	-3.59**	-1.54	NA	-3.33**	-3.61**
		(.42)	(.81)	(2.21)		(.57)	(1.12)
Classroom % Minority	-7.68**	-.18**	-3.70**	-1.82**	-6.46**	-2.45**	-.82*
	(2.14)	(.03)	(.32)	(.42)	(1.89)	(.28)	(.36)
Missing Minority Percent	NA	-1.33**	-.85^{m}	.09	NA	-.43**	-1.19**
		(.34)	(.45)	(.72)		(.11)	(.13)
Ability Rank (low as reference)							
Rank 2	NA	.88**	1.57**	2.41^{m}	NA	1.46**	.1.44**
		(.14)	(.24)	(1.27)		(.14)	(.16)
Rank 3	1.06	2.11**	2.65**	6.16**	3.75*	2.56**	3.18**
	(.87)	(.14)	(.23)	(1.19)	(1.67)	(.14)	(.16)
Rank 4	3.85**	3.00*	4.37**	5.36**	7.52**	4.04**	5.16**
	(1.17)	(.13)	(.23)	(1.23)	(1.66)	(.14)	(.16)
Rank 5 (highest)	5.67**	5.02**	6.55**	10.04**	10.64**	6.59**	7.99**
	(1.28)	(.14)	(.24)	(1.17)	(1.66)	(.14)	(.16)
Interaction of Ability Rank & Hands-on Scale (Scale X Low Rank as reference)							
Scale X Ability Rank 2				-.58	NA		
				(.46)			
Scale X Ability Rank 3				-1.03*	-.88^{m}		
				(.43)	(.48)		
Scale X Ability Rank 4				-.20	-1.14*		
				(.44)	(.50)		
Scale X Ability Rank 5				-.85*	-1.50**		
				(.42)	(.44)		
n	1231	10,072	5435	3112	1231	15,241	10,438
R^2	.29	.23	.29	.31	.31	.25	.35

Table 5-6: Cross-Sectional NELS Model for Teacher Hands-on Scale and Test Score from NELS Data (1988, 1990, 1992)

Panel I

	Non-Imputed Data			Imputed Data		
Variable	**8th Grade**	**10th Grade**	**12th Grade**	**8th Grade**	**10th Grade**	**12th Grade**
Teacher reported hands-on science scale	.45**	.38**	-.10	.43**	.32**	-.02
	(.07)	(.11)	(.14)	(.07)	(.10)	(.12)
Female	-1.17**	-2.31**	-2.08**	-1.16**	-2.32**	-1.89**
	(.09)	(.14)	(.18)	(.08)	(.13)	(.17)
Race/Ethnicity (White as reference)						
Asian	-.44*	.01	-1.03**	-.35^{m}	.02	-.89**
	(.19)	(.31)	(.34)	(.18)	(.28)	(.32)
Black	-2.04**	-1.79**	-2.33**	-2.00**	-1.91**	-2.26**
	(.16)	(.29)	(.38)	(.15)	(.26)	(.36)
Hispanic	-.94**	-1.02**	-1.38**	-.90**	-1.05**	-1.39**
	(.16)	(.27)	(.38)	(.15)	(.25)	(.34)
Race/Ethnicity Missing				-2.18**	-1.21	-2.02
				(.39)	(.96)	(2.29)
Ability Rank Quintiles (Rank 1, low, as reference)						
Ability Rank 2	.61**	1.01**	-.24	.50**	1.00*	-.33
	(.15)	(.25)	(.35)	(.13)	(.22)	(.30)
Ability Rank 3	1.46**	1.73**	1.93**	1.31**	1.54**	1.03**
	(.14)	(.24)	(.34)	(.13)	(.21)	(.31)
Ability Rank 4	2.06**	2.68**	1.48**	1.85**	2.76**	1.43**
	(.14)	(.24)	(.35)	(.13)	(.22)	(.32)
Ability Rank 5	3.59**	4.55**	3.35**	3.45**	4.42**	3.53**
	(.15)	(.25)	(.37)	(.14)	(.23)	(.34)
Family SES	1.40**	1.53**	1.33**	1.40**	1.50**	1.31**
	(.07)	(.11)	(.13)	(.06)	(.10)	(.12)
Hours of Science Homework	.33**	..07	-.02	.31**	.07	.08
	(.07)	(.08)	(.10)	(.06)	(.08)	(.09)
Missing Hours of Science Homework				-.18	-.98**	-2.21**
				(.22)	(.37)	(.60)
Science Course Taken (versus other science courses)	NA			NA		
Biology		.88*	.42		.85**	.70
		(.27)	(.55)		(.25)	(.47)
Missing Biology					.78	NA
					(.66)	
Chemistry		1.63**	1.32**		1.57**	1.33**
		(.21)	(.27)		(.20)	(.24)
Missing Chemistry					-.58	NA
					(.58)	

table 5-6 continued

Table 5-6

Panel II

	Non-Imputed Data			Imputed Data		
Variable	**8th**	**10th**	**12th**	**8th**	**10th**	**12th**
	Grade	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**
Earth Science		.15	.19		.09	.36
		(.18)	(.22)		(.17)	(.21)
Missing Earth Science					.22	NA
					(.47)	
Physics		-.10**	1.54**		-.99**	1.58**
		(.37)	(.21)		(.34)	(.20)
Missing Physics					.34	NA
					(.62)	
Number of Courses Taken	NA	-.20**	.44**	NA	-.17**	.45**
		(.06)	(.12)		(.05)	(.11)
Missing Number of Courses Taken	NA			NA	NA	NA
Classroom Level Variables						
School/Class Minority Percent	-.14**	-2.46**	-1.30**	-.14**	-2.31**	-1.46**
	(.03)	(.36)	(.46)	(.03)	(.32)	(.42)
Missing Minority Percent				-.85**	-.93*	-.34
				(.31)	(.42)	(.67)
Science Class Track						
(Other Track as reference)						
High Track	2.79**	.75**	.15	.82**	.80**	.15
	(.11)	(.17)	(.27)	(.11)	(.15)	(.25)
Missing Track				.40**	NA	.49
				(.16)		(.89)
Achievement Level of Students in Class						
(Low Achievement as reference)						
Differential Achievement	1.07**	1.64**	.81	1.06**	1.55**	.51
	(.16)	(.30)	(.49)	(.14)	(.27)	(.42)
Average Achievement	.85**	1.55**	$.79^{m}$	.84**	1.46**	.61*
	(.14)	(.23)	(.42)	(.13)	(.21)	(.36)
High Achievement	2.31**	2.67**	1.96**	2.28**	2.65**	1.66**
	(.16)	(.27)	(.46)	(.15)	(.25)	(.40)
Missing Achievement				2.15**	.95*	2.78*
				(.61)	(.47)	(1.11)

table 5-6 continued

Table 5-6

Panel III

	Non-Imputed Data			Imputed Data		
Variable	**8th Grade**	**10th Grade**	**12th Grade**	**8th Grade**	**10th Grade**	**12th Grade**
School Level Variables						
School Control (Public as reference)						
Catholic	-.25	-.12	-.12	-.33*	.11	-.31
	(.16)	(.34)	(.36)	(.15)	(.31)	(.34)
Private - other religious	-.18	-.001	.19	-.14	.10	-.04
	(.23)	(.53)	(.62)	(.21)	(.51)	(.55)
Private - non-religious	1.55**	2.00**	.50	1.58**	2.29**	.64^{m}
	(.19)	(.34)	(.38)	(.18)	(.31)	(.36)
Metropolitan Status (Urban as reference)						
Suburban	-.13	-.10	-.75**	-.14	.04	-.79**
	(.11)	(.20)	(.24)	(.10)	(.20)	(.23)
Rural	-.03	-.10	.44**	-.10	.08	-.57*
	(.13)	(.22)	(.27)	(.12)	(.20)	(.25)
Geographic Location (South as reference)						
Northeast	.02	.51*	..05	-.01	.45*	.11
	(.13)	(.22)	(.26)	(.12)	(.20)	(.25)
North central	.21	.66**	.58*	.23*	.55**	.48*
	(.11)	(.18)	(.24)	(.10)	(.17)	(.22)
West	.06	.68**	.94**	.12	.56*	.81**
	(.13)	(.21)	(.28)	(.12)	(.19)	(.26)
N	8737	4443	2630	10071	5283	3102
R^2	.34	.40	.40	.34	.41	.43

Table 5-7: Cross-Sectional NELS Model of Science Test Scores and Student Hands-on Scale Using 10th Grade NELS Data (1990)

Panel I

Variable	Non-Imputed		Imputed	
	10th Grade (w/o teacher reported classroom variables)	10th Grade (with teacher reported classroom variables)	10th Grade (w/o teacher reported classroom variables)	10th Grade (with teacher reported classroom variables)
Individual Level Variable				
Student reported hands-on science scale	-.06	-.10	-.05	.05
	(.05)	(.08)	(.05)	(.05)
Female	-2.21**	-2.32**	-2.07**	-2.16**
	(.08)	(.14)	(.08)	(.08)
Race/Ethnicity (White as reference)				
Asian	-.44*	.05	-.38*	-.41**
	(.18)	(.31)	(.23)	(.17)
Black	-3.08**	-1.80**	-3.08**	-2.92**
	(.16)	(.29)	(.15)	(.15)
Hispanic	-1.91**	-.88**	-1.83**	-1.73**
	(.15)	(.27)	(.14)	(.14)
Missing Race/Ethnicity			-3.48**	-2.69**
			(.85)	(.85)
Ability Rank Quintiles (Rank 1 as reference)				
Ability Rank 2	.92**	1.10**	.91**	.88**
	(.14)	(.25)	(.13)	(.13)
Ability Rank 3	1.68**	1.80**	1.61**	1.54**
	(.14)	(.24)	(.13)	(.13)
Ability Rank 4	2.71**	2.76**	2.73**	2.61**
	(.14)	(.24)	(.13)	(.13)
Ability Rank 5	4.91**	4.63**	4.80**	4.61**
	(.15)	(.25)	(.14)	(.14)
Family SES	1.84**	1.55**	1.84**	1.76**
	(.06)	(.11)	(.06)	(.06)
Hours of Science Homework	.27**	.11	.28**	.24**
	(.04)	(.08)	(.05)	(.05)
Missing Hours of Science Homework			-1.89**	-1.80**
			(.23)	(.23)
Student answered "not in science class" in response To Hours of Science Homework item	.31	NA	.33	-.15
	(.22)		(.25)	(.21)

table 5-7 continued

Table 5-7

Panel II

Variable	Non-imputed w/o classroom	Non-imputed w classroom	Imputed w/o classroom	Imputed w classroom
Science Courses Taken (vs. other science courses)				
Biology	1.76** (.15)	.87** (.26)	1.78** (.14)	1.61** (.14)
Missing Biology			.38 (.36)	.34 (.36)
Chemistry	2.13** (.12)	1.70** (.21)	2.16** (.12)	2.02** (.12)
Missing Chemistry			.02 (.35)	-.02 (.35)
Earth Science	.10 (.11)	.19 (.18)	.11 (.10)	.12 (.10)
Missing Earth Science			$-.51^{m}$ (.27)	$-.47^{m}$ (.27)
Physics	-.80** (.22)	-.98** (.38)	-.73** (.21)	-.68** (.21)
Missing Physics			-.42 (.37)	.40 (.36)
Number of Science Courses Taken	-.27** (.03)	-.22** (.06)	-.28** (.03)	-.27** (.03)
Missing Number of Science Courses Taken			NA	NA
Class Level Variables				
School/Class Minority Percent		-2.42** (.36)		-.96** (.27)
Missing Minority Percent				-.13 (.32)
Science Class Track (Low Track as reference)				
High Track		.84** (.17)		.60** (.15)
Missing Track				NA
Achievement Level of Students in Class (Low Achievement as reference)				
Differential Achievement		1.62** (.30)		1.32** (.28)
Average Achievement		1.54** (.23)		1.27** (.21)
High Achievement		2.64** (.27)		2.31** (.24)
Missing Achievement				1.86** (.36)

table 5-7 continued

Table 5-7

Panel III Variable	Non-imputed w/o classroom	Non-imputed w classroom	Imputed w/o classroom	Imputed w classroom
School Level Variables				
School Control				
(Public as reference)				
Catholic	-.04	.0001	.01	-.01
	(.19)	(.37)	(.18)	(.18)
Private - other religious	.07	-.16	.23	.27
	(.26)	(.54)	(.25)	(.25)
Private - non-religious	2.16**	2.02**	2.27**	2.30**
	(.20)	(.34)	(.19)	(.19)
Metropolitan Status				
(Urban as reference)				
Suburban	-.01	-.12	.07	.05
	(.11)	(.21)	(.11)	(.11)
Rural	.27	-.17	.35*	.31*
	(.12)	(.22)	(.12)	(.11)
Geographic Location				
(South as reference)				
Northeast	.90**	.47*	.91**	.88**
	(.12)	(.22)	(.12)	(.12)
North central	.94**	.74**	.90**	.88**
	(.11)	(.19)	(.10)	(.11)
West	.40**	.74**	.44**	.46**
	(.13)	(.21)	(.12)	(.12)
N	13,001	4413	14,715	14,715
R^2	.36	.40	.37	.38

Table 5-8: Cross-Sectional NELS Model of Science Test Scores and Student Hands-on Scale Using 12th Grade NELS Data (1992)

Panel I

Variable	**Non-Imputed**		**Imputed**	
	12th Grade (w/o teacher reported classroom variables)	**12th Grade (with teacher reported classroom variables)**	**12th Grade (w/o teacher reported classroom variables)**	**12th Grade (with teacher reported classroom variables)**
Individual Level Variable				
Student reported hands-on science scale	-.52** (.06)	-.59** (.12)	-.52** (.06)	-.52** (.06)
Female	-2.34** (.9)	-2.15** (.08)	-2.26** (.09)	-2.27** (.09)
Race/Ethnicity (White as reference)				
Asian	-1.45** (.18)	-1.00** (.33)	-1.44** (.18)	-1.44** (.18)
Black	-3.21** (.18)	-2.23** (.38)	-3.17** (.17)	-3.14** (.17)
Hispanic	-1.71** (.16)	-1.43** (.38)	-1.67** (.16)	-1.67** (.16)
Missing Race/Ethnicity			-3.02* (1.27)	-3.15* (1.27)
Ability Rank Quintiles (Rank 1 as reference)				
Ability Rank 2	.64** (.16)	-.07** (.34)	.52** (.15)	.51** (.15)
Ability Rank 3	1.39** (.16)	1.01** (.34)	1.32** (.16)	1.29** (.16)
Ability Rank 4	2.36** (.17)	1.52** (.36)	2.26** (.16)	2.22** (.16)
Ability Rank 5	4.20** (.18)	3.33** (.37)	4.15** (.17)	4.00** (.18)
Family SES	1.37** (.07)	1.29** (.23)	1.39** (.07)	1.37** (.07)
Hours of Science Homework	.22** (.06)	.07 (.10)	.26** (.06)	.19** (.06)
Missing Hours of Science Homework			-1.85** (.35)	-1.82** (.37)
Student answered "not in science class" in response To Hours of Science Homework item	.28^{m} (.15)	-.66 (1.50)	.26^{m} (.14)	.28^{m} (.15)

table 5-8 continued

Table 5-8

Panel II

Variable	Non-imputed w/o classroom	Non-imputed w classroom	Imputed w/o classroom	Imputed w classroom
Science Courses Taken (vs. other science courses)				
Biology	-.07	.11	-.01	-.02
	(.23)	(.56)	(.21)	(.23)
Chemistry	1.79**	1.28**	1.80**	1.77**
	(.13)	(.27)	(.12)	(.12)
Earth Science	.26*	.24	.29*	.31**
	(.12)	(.22)	(.11)	(.11)
Physics	1.84**	1.59**	1.85**	1.78**
	(.13)	(.21)	(.12)	(.12)
Number of Science Courses Taken	.57**	.42**	.57**	.54**
	(.07)	(.12)	(.07)	(.07)
Class Level Variables				
School/Class Minority Percent		-1.22**		-.98**
		(.45)		(.24)
Missing Minority Percent				.13
				(.42)
Science Class Track (Other Track as reference)				
High Track		.19		-.04**
		(.27)		(.24)
Missing Track				-.03
				(.72)
Achievement Level of Students in Class (Low Achievement as reference)				
Differential Achievement		.67		.24
		(.49)		(.43)
Average Achievement		.74^{m}		.44
		(.42)		(.37)
High Achievement		1.97**		1.23**
		(.46)		(.39)
Missing Achievement				.41
				(.82)

table 5-8 continued

Table 5-8

Panel III

	Non-imputed w/o classroom	Non-imputed w classroom	Imputed w/o classroom	Imputed w classroom
School Level Variables				
School Control				
(Public as reference)				
Catholic	-.42*	-.12	-.43*	-.39*
	(.20)	(.36)	(.20)	(.20)
Private - other religious	.52^{m}	.29	.51^{m}	.51*
	(.30)	(.60)	(.28)	(.28)
Private - non-religious	.38^{m}	.63	.41**	.46*
	(.21)	(.38)	(.20)	(.20)
Metropolitan Status				
(Urban as reference)				
Suburban	-.02	-.70**	-.02	-.03
	(.12)	(.24)	(.12)	(.12)
Rural	.10	-.43	.11	.08
	(.13)	(.27)	(.13)	(.13)
Geographic Location				
(South as reference)				
Northeast	.44**	.12	.47**	.51**
	(.14)	(.26)	(.13)	(.14)
North central	.52**	.67**	.52**	.52**
	(.12)	(.23)	(.12)	(.12)
West	.48**	1.01**	.50**	.48**
	(.14)	(.28)	(.14)	(.14)
N	9,895	2,628	10,389	10,389
R^{2}	.46	.41	.46	.46

Table 5-9: Coefficients on Teacher Hands-on Scale

Scale	8th Grade	10th Grade	12th Grade
Original Scale	.43** (.07)	.32** (.10)	-.02 (.12)
1. Time on Lab Periods	.20** (.04)	.14^{m} (.08)	-.07 (.09)
2. Student Experiments	.18** (.04)	.31** (.08)	.04 (.12)
3. Time on Lab Periods + Student Experiments	.23** (.04)	.33** (.09)	-.05 (.11)
4. Student Experiments + Student Reports	NA	.32** (.08)	.05 (.13)
5. Amount and Condition of Scientific Equipment	.33** (.05)	.33** (.11)	.12 (.13)

Table 5-10: Coefficients on Course Taking Variables

Variable	10th Grade		12th Grade	
	Teacher Scale	Student Scale	Teacher Scale	Student Scale
Biology	.85** (.25)	1.61** (.14)	.70 (.47)	-.02 (.23)
Chemistry	1.57** (.20)	2.02** (.12)	1.33** (.24)	1.77** (.12)
Earth Science	..09 (.17)	.12 (.10)	.36 (.21)	.31** (.11)
Physics	-.99* (.34)	-.68** (.21)	1.58** (.20)	1.78** (.12)
Total Courses	-.17** (.05)	-.27** (.03)	.45** (.11)	.54** (.07)

Table 5-11: Top 5 Patterns of Course Taking for 10th Grade Students

Course Pattern	**Percent of Whole Sample**	**N**
Biology (B)	9	1,456
Biology, Chemistry (B,C)	9	1,584
Biology, Earth Science (B,E)	10	1,788
Biology, Other Science (B,O)	33	5,755
Biology, Earth, Other (B,E,O)	7	1,135

Table 5-12: Mean of Hands-On Science Scale By Course Pattern: 10th Grade
(Significant differences between groups shown in superscript)

Index	Course Pattern	Teacher Scale	N	Student Scale	N
	Full Sample	2.74 (.70)	5,400	2.48 (.90)	15,200
1	B	2.76^{2} (.64)	493	2.47^{2} (.89)	1,433
2	B,C	$2.90^{1,3\text{-}5}$ (.62)	549	$2.77^{1,3\text{-}5}$ (.73)	1,583
3	B,E	2.76^{2} (.66)	620	$2.49^{2,4}$ (.85)	1,780
4	B,O	2.70^{2} (.70)	2,201	$2.41^{2,3,5}$ (.88)	5,709
5	B,E,O	2.69^{2} (.65)	372	$2.54^{2,4}$ (.92)	1,119
	Top 5 Patterns Together	2.74 (.68)	4,235	2.49 (.87)	11,624

Table 5-13: Coefficient on Hands-on Scale By Course Pattern: 10th Grade

Course Pattern	Teacher Scale	Student Scale (with class level covariates)	Student Scale (without class level covariates)
Whole Sample	.32** (.10)	-.05 (.46)	-.05 (.46)
B	**-.04** (.35)	.09 (.15)	.07 (.15)
B,C	.53^{m} (.31)	-.14 (.16)	-.15 (.16)
B,E	**.84**** (.31)	**.23** (.14)	**.26^{m}** (.14)
B,O	.34* (.15)	**-.24**** (.08)	**-.24**** (.08)
B,E,O	.43 (.41)	.001 (.17)	.03 (.17)

Table 5-14: Results from Extended NELS Model Including Course Pattern and Interaction Term for Teacher Hands-on Scale: 10th Grade

Variable	Original Model	Interaction of B & Scale	Interaction of B,E & Scale
Teacher Scale	.32** (.10)	.34** (.11)	.23* (.11)
B		.56 (1.16)	-.20 (.64)
Interaction of B*Teacher Scale		-.28 (.35)	NA
B,E		2.08** (.65)	-.09 (1.06)
Interaction of B,E*Teacher Scale		NA	.79* (.31)
B,C		.79^{m} (.46)	.78^{m} (.46)
B,O		.86 (.59)	.87 (.59)
B,E,O		.68 (.66)	.70 (.66)
R^2	.40	.42	.42

Table 5-15: Results from Extended NELS Model Including Course Pattern and Interaction Term for Student Hands-on Scale: 10th Grade

Variable	Original Model		Including Interaction of B,E & Scale		Including Interaction of B,O & Scale	
	With Class Level Covariates	w/o Class-Level Covariates	With Class Level Covariates	w/o Class-Level Covariates	With Class Level Covariates	w/o Class-Level Covariates
Student Scale	-.05 (.05)	-.05 (.05)	-.07 (.07)	-.07 (.05)	$.11^{m}$ (.06)	$.11^{m}$ (.06)
B,E			2.23** (.53)	2.04** (.52)	2.63** (.38)	2.86** (.38)
Interaction of B,E & Scale			.17 (.15)	.34* (.15)	NA	NA
B,O			2.29** (.42)	1.64** (.35)	2.37** (.42)	2.52** (.42)
Interaction of B,O & Scale			NA	NA	-.35** (.10)	-.37** (.10)
B			.74 (.38)	.93* (.38)	.74* (.38)	.92* (.38)
B,C			1.38** (.27)	1.48** (.27)	1.37** (.27)	1.47** (.27)
B,E,O			1.22** (.38)	1.46** (.38)	1.21** (.38)	1.42** (.38)
R^2	.38	.37	.39	.38	.39	.38

Table 5-16: Top 5 Patterns of Course Taking for 12th Grade Students

Course Pattern	Percent of Whole Sample	N
Biology, Chemistry (B,C)	7	1,232
Biology, Other (B,O)	20	3,319
Biology, Chemistry, Other (B,C,O)	14	1,655
Biology, Chemistry, Physics (B,C,P)	10	2,411
Biology, Chemistry, Other, Physics (B,C,O,P)	8	1,354

Table 5-17: Mean of Hands-On Science Scale By Course Pattern: 12th Grade
(Significant differences between groups shown in superscript)

Index	Course Pattern	Teacher Scale	N	Student Scale	N
	Whole Sample	2.73 (.65)	3,820	2.77 (.80)	13,382
1	B,C	2.76 (.66)	277	2.79^{3} (.76)	1,092
2	B,O	2.68 (.77)	406	2.74 (.85)	2,164
3	B,C,P	2.69 (.65)	628	$2.70^{1,4}$ (.71)	1,559
4	B,C,O	2.79 (.63)	634	$2.80^{3,5}$ (.79)	2,227
5	B,C,P,O	2.76 (.63)	635	2.70^{4} (.75)	1,291
	Total for 5 Patterns	2.74 (.67)	2,580	2.75 (.78)	8,333

Table 5-18: Coefficient on Hands-on Scale for Sub-Samples of Course-Taking: 12th Grade

Course Pattern	**Teacher Scale**	**Student Scale (with class level covariates)**	**Student Scale (without class level covariates)**
Whole Sample	.10	-.46**	-.45**
B,C	**-.20**	-.76**	-.79**
B,O	-.04	-.29*	-.26*
B,C,P	.01	-.76**	-.80**
B,C,O	.24	-.65**	-.67**
B,C,P,O	**.57**^{m}	-.52**	-.49**

Table 5-19: Results from Extended NELS Model Including Course Pattern and Interaction Term Teacher Hands-on Scale: 12th Grade

Variable	Original Model	Including B,C	Including B,C,P,O
Teacher Scale	.10 (.13)	.004 (.13)	-.11 (.14)
B,C		.24 (1.40)	-.72 (.62)
Interaction of B,C & Scale		-.34 (.46)	NA
B,C,O,P		-.69 (.52)	-2.21* (1.09)
Interaction of B,C,O,P & Scale		NA	.55 (.34)
B,O		-.49 (.49)	-.47 (.49)
B,C,P		-.42 (.52)	-.44 (.52)
B,C,O		-1.45* (.49)	-1.44* (.59)
R^2	.40	.42	.42

Appendix Tables

Table A5-1: NELS 8th Grade: A Comparison of the Full Sample and the Sub-Samples Available for Analysis (unweighted)

Variable	NELS 8th Grade Sample	Sample with Science Test Score	Sample Available for Analysis of Teacher Scale
n (based on the number of student id)	24,599	23,616	10,221
	%	%	%
Gender			
Female	50.2	50.3	50.4
Male	49.8	49.8	49.6
Race/Ethnicity			
Asian	6.4	6.4	5.9
Black	12.5	12.3	11.2
Hispanic	13.2	13.0	11.9
White	67.9	68.3	70.1
SES Quartiles (1 is low)			
Quartile 1	24.2	23.9	22.3
Quartile 2	23.4	23.4	23.2
Quartile 3	23.7	23.8	23.9
Quartile 4	28.7	28.9	30.7
Ability Rank Quintiles (1 is low)			
Ability Rank 1	19.9	19.7	19.4
Ability Rank 2	16.8	16.7	17.0
Ability Rank 3	22.5	22.5	21.8
Ability Rank 4	23.4	23.5	23.6
Ability Rank 5	17.5	17.7	18.2
School Control			
Public	78.9	78.6	78.6
Catholic	10.6	10.7	10.1
Private School - other religion	4.4	4.5	4.4
Private School – non-religious	6.1	6.1	6.9
Geographic Region			
Northeast	20.0	20.0	16.9
North Central	25.0	25.3	26.8
South	34.5	34.5	37.0
West	20.5	20.3	29.3
Metro Status			
Urban	31.0	30.7	29.3
Suburban	41.7	41.7	42.5
Rural	27.4	27.6	28.2

table A5-1 continued

Table A5-1

Variable	NELS 8th Grade Sample	Sample with Science Test Score	Sample Available for Analysis of Teacher Scale
School Minority Percent			
0. 0%	12.3	12.4	10.5
1. 1 - 5%	21.8	22.2	22.2
2. 6 - 10%	11.1	11.1	13.1
3. 11 - 20%	13.2	13.4	14.8
4. 21 - 40%	15.4	15.0	16.1
5. 41 - 60%	9.0	9.0	9.9
6. 61 - 90%	9.4	9.3	7.2
7. 91 - 100%	7.8	7.6	6.1

Table A5-2: NELS 10th Grade: A Comparison of the Full Sample and the Sub-Samples Available for Analysis (unweighted)

Variable	NELS 10th Grade Sample	Sample With Science Test Score	Sample Available for Analysis of Student Scale	Sample Available for Analysis of Teacher Scale
N	18,104	16542	15,478	5,449
(based on number of student id)				
	%	%	%	%
Gender				
Female	50	50.3	51	52
Male	50	49.7	49	48
Race/Ethnicity				
Asian	6.8	6.6	6.7	6.1
Black	10.2	9.8	9.4	9.1
Hispanic	12.7	11.9	11.3	10.9
White	69.4	71.8	72.7	73.9
SES Quartiles				
(1 is low)				
Quartile 1	21.7	21.2	20.2	19.0
Quartile 2	24	24.0	23.6	23.7
Quartile 3	24	24.3	24.5	25.2
Quartile 4	30.3	30.5	31.8	32.1
Ability Rank Quintiles				
(1 is low)				
Ability Rank 1 (low)	18.7	18.4	17.5	16.0
Ability Rank 2	18.2	17.9	17.6	17.3
Ability Rank 3	21.9	22.0	22.1	22.9
Ability Rank 4	22.3	22.4	23.0	23.0
Ability Rank 5	18.9	19.2	19.8	20.7
School Control				
Public	86.4	86.3	85.8	86.9
Catholic	5.6	5.8	6.0	5.4
Private School - other religion	2.6	2.7	2.7	1.6
Private School –non-religious	5.4	5.2	5.5	6.2
Geographic Region				
Northeast	18.8	18.7	19.1	16.9
North Central	26.2	27.1	27.0	25.8
South	34.5	35.0	35.3	39.4
West	20.6	19.3	18.7	18.0
Metro Status				
Urban	29.5	28.7	28.5	27.3
Suburban	39.6	39.6	40.1	40.2
Rural	30.1	31.7	31.5	32.5

table A5-2 continued

Table A5-2

Variable	NELS 10th Grade Sample	Sample With Science Test Score	Sample Available for Analysis of Student Scale	Sample Available for Analysis of Teacher Scale
	%	%	%	%
School Minority Percent				
0. 0%	24.3	24.5	25.1	25.1
1. 1 - 5%	8.0	8.1	8.1	8.1
2. 6 - 10%	11.7	11.7	11.9	11.9
3. 11 - 20%	18.1	18.0	18.3	18.3
4. 21 - 40%	15.9	16.0	16.1	16.1
5. 41 - 60%	8.7	8.7	8.6	8.6
6. 61 - 90%	7.2	7.1	6.6	6.6
7. 91 - 100%	6.1	5.9	5.2	5.2

Table A5-3: NELS 12th Grade: A Comparison of the Full Sample and the Sub-Samples Available for Analysis (unweighted)

Variable	NELS 12th Grade	Sample With Science Test Score	Sample Available for Analysis of Student Scale	Sample Available for Analysis of Teacher Scale
N	16,977	12,734	11,203	3,180
(based on number of student id)				
Gender				
Female	50	50.2	51	51
Male	50	49.8	49	49
Race/Ethnicity				
Asian	7.4	7.1	7.7	8.7
Black	9.5	9.2	9.0	9.2
Hispanic	12.0	11.7	11.8	7.1
White	71.1	72.0	71.5	75.0
SES Quartiles				
(1 is low)				
Quartile 1	19.0	19.0	17.2	13.3
Quartile 2	23.3	23.6	22.6	20.1
Quartile 3	25.2	25.2	25.6	26.5
Quartile 4	32.6	32.2	34.6	40.1
Ability Rank Quintiles				
(1 is low)				
Ability Rank 1	20.0	19.1	17.3	13.3
Ability Rank 2	20.3	20.2	19.2	16.6
Ability Rank 3	19.7	20.2	20.0	20.5
Ability Rank 4	20.5	20.8	21.9	23.9
Ability Rank 5	19.5	19.7	21.6	25.7
School Control				
Public	86.1	85.6	84.3	82.8
Catholic	5.5	6.0	6.5	7.3
Private School - other religion	2.6	2.7	2.9	2.7
Private School –non-religious	5.8	5.7	6.3	7.2
Geographic Region				
Northeast	19.6	18.9	19.8	20.0
North Central	26.5	27.2	25.7	27.9
South	33.7	35.0	34.9	33.2
West	20.2	19.0	19.5	19.0
Metro Status				
Urban	28.9	28.1	29.5	29.4
Suburban	40.7	38.9	39.4	40.0
Rural	30.5	33.0	31.2	31.0

table A5-3 continued

Table A5-3

Variable	NELS 12th Grade	Sample With Science Test Score	Sample Available for Analysis of Student Scale	Sample Available for Analysis of Teacher Scale
	%	%	%	%
School Minority Percent				
0. 0%	21.7	22.9	22.6	22.4
1. 1 - 5%	8.7	8.9	9.0	9.5
2. 6 - 10%	13.3	13.4	13.4	13.2
3. 11 - 20%	18.3	17.9	18.2	19.5
4. 21 - 40%	16.7	16.3	16.4	16.4
5. 41 - 60%	7.5	7.4	7.4	7.6
6. 61 - 90%	7.5	7.0	6.8	6.7
7. 91 - 100%	6.4	6.3	6.1	4.7

Chapter 6: Conclusion

Revisiting Our Objectives

We set out with this work to examine the relationship of hands-on science with student achievement. Our work arises from the current intersection of two science education policies. The first is the ongoing promotion of hands-on science for the past 40 years and the second a decade-long science reform effort that would reduce the use of hands-on science. Our concern arises from the mixed results on past research regarding the relationship between hands-on science and student achievement. Without a conclusive finding on this relationship, it is not clear which of the two is the better policy to implement.

Our fundamental research question is whether there is a positive link between hands-on science and student achievement. Past research on this topic has examined the relationship of the level of hands-on science in the classroom with student scores on standardized tests. We have identified four issues that may contribute to the inconclusiveness of this past work as well as have implications for current policy. These issues include: 1) the need to control for variables that may be linked to both hands-on science and test scores, 2) the need to examine the link between hands-on science and performance test scores as well as the more traditional multiple choice test scores - the importance of which has increased with the current debate over the adequacy of using multiple choice tests to measure student achievement, 3) the need to investigate a potential differential relationship of hands-on science and test score due to student ability, and 4) the need to consider the multiple facets of hands-on science including the quantity, quality and instructional approach used with it.

With guidance from the theoretical and empirical literature, we developed three hypotheses from our research question and the above four issues. Specifically, we hypothesized that:

1. Higher levels of hands-on science are associated with higher test scores, be they multiple choice or performance tests, all else being equal.
2. This association is stronger with performance tests than with multiple choice tests.
3. This association is weaker for higher ability students.

To test these hypotheses, we analyzed two data sets allowing us to perform complimentary analyses, giving us greater confidence in our results, as well as addressing the first three issues. Unfortunately, we were not able to address the fourth issue as data limitations did not allow us to discern the quality of nor the instructional approach used with the hands-on science. Our primary data set, the RAND data, has a large 8th grade student sample from the Southern California region but a small teacher sample. It is used to test all three hypotheses. The advantage of the RAND data is that it contains both multiple choice and performance test scores for the same students. A disadvantage is the lack of some covariates we would like to see included. The NELS is nationally representative and contains the covariates linked to the level of hands-on science and student test scores. It has both a large student and teacher sample but lacks performance test data and an 8th grade student survey of hands-on science. We use it to test Hypotheses 1 and 3. NELS also contains data on students in 10th and 12th grades allowing us to examine the relationship in higher grades as to its similarity or difference (as predicted by developmental theory) with 8th grade results.

Overall Findings

The table below lays out the overall findings from the analysis. Column 1 lists the survey and the grade level of the sample. Column 2 identifies the source of the hands-on science measure: reported by student or teacher. Columns 3 & 4 are concerned with the evidence for Hypothesis 1 and note whether positive relationships between the hands-on science scale and the multiple choice test scores or performance test scores were found. Column 5 lists whether evidence was found in support of Hypothesis 2 that the relationship would be stronger with performance test scores. Columns 6 & 7 concern Hypothesis 3 and show whether higher ability students have a less positive relationship between hands-on work and multiple choice or performance test scores. These results should be considered with the strengths and weaknesses of the two data sets kept in mind (a topic we will return to in the next section).

Table 6.1: Overall Results

Survey	**Hands-on Measure**	**Hypothesis 1**		**Hypothesis 2**	**Hypothesis 3**	
		MC Test	PA Test		MC Test	PA Test
RAND						
8th Grade	Student	Yes[1]	Yes	Yes[2]	Yes	No/Yes[3]
	Teacher	No	No	No	No	No
NELS						
8th Grade	Teacher	Yes	NA	NA	No	NA
10th Grade	Student	No[4]	NA	NA	No	NA
10th Grade	Teacher	Yes	NA	NA	No	NA
12th Grade	Student	No	NA	NA	No	NA
12th Grade	Teacher	No	NA	NA	No	NA

[1] For students of lower Ability Ranks [2] For students of higher Ability Ranks

[3] Not found with full scale but found when using between-class variation in the scale.

[4] Incomplete scale

Let us start by summarizing the evidence regarding Hypothesis 1 using the 8th grade results as these are the focus of our primary data set. The RAND results support Hypothesis 1 in regards to students of lower Ability Ranks when using the student reports but provide no supporting evidence when using the teacher data. The NELS 8th grade results, based on teacher reports and multiple choice test scores, also support Hypothesis 1 but without regard to student Ability Rank. As the NELS 8th grade student survey did not include items on hands-on science, we have no results using a student scale. Hence we are faced with the seemingly inconsistent result that teacher reports from one data set support our hypothesis while teacher reports from another do not. This inconsistency is not a total surprise for the RAND data has a low variation in the teacher scale due to the small number of teachers taking part making it difficult to find a relationship.

Looking to the NELS results using higher grades for further evidence leads to further mixed evidence. The NELS teacher reports support the hypothesis for 10th grade students as they did for 8th grade. However, the 10th grade student reports do not and we again face inconsistent results when using reports from the same source (this time students) in different data sets. Here again, this inconsistency is not unexpected as the NELS 10th grade student scale is incomplete and may not be an adequate measure of hands-on science. For the 12th grade results, the results based on either the student and teacher data do not support Hypothesis 1. Further consideration must be given to the appropriateness of comparing 12th graders with 8th graders before making conclusions based on this finding.

Concerning Hypothesis 2, we can draw upon our RAND results which show a stronger relationship for hands-on science with one type of performance test than with the multiple choice test for higher ability students but no difference for lower ability students. Our finding for higher ability students is not really in the spirit of Hypothesis 2 which was based on the idea that hands-on science would better prepare students for performance tests. Instead, our finding occurs because for higher ability students hands-on science is negatively linked to multiple choice test scores rather than associated with higher performance test scores.

For Hypothesis 3, the RAND student results provide evidence that hands-on science is positively linked to test scores for lower ability students for both types of test while this relationship does not occur for higher ability students. For performance test scores, we see these findings after breaking out the between-class variation in the scale. However, the NELS results show a uniform positive relationship between hands-on science and test score, when using the 8th and 10th grade teacher data, for all students regardless of student ability. Because the NELS model includes additional relevant covariates we have greater confidence in them and overall we do not find enough evidence to support Hypothesis 3.

Findings for the major covariates are similar for all the analyses. We see positive relationships between test scores and higher SES and higher student ability (individually in both data sets as well as in high class track and high class achievement level in the NELS data), and negative relationships between test scores and Hispanic, Black and classroom percentage minority.

The findings were further examined through a series of sensitivity analyses which confirmed their robustness. For the RAND data, we broke the student scale down into its between-class and within-class variation as the former may be a better measure of the actual differences in hands-on science. Additionally we checked if the results were maintained for different student characteristics and considered the need for non-linearities in the relationship of hands-on science to test score. For the NELS data, we examined alternative teacher scales (based on uniformity between the three waves). In addition we performed a detailed analysis of the impact of course taking on the relationship including current enrollment in a science course and the pattern of courses taken.

In conclusion, we find little evidence to support Hypotheses 2 & 3. After a first glance at the data for Hypothesis 1, the results from the two data sets appear contradictory. The analysis of the RAND data found a relationship between test scores and the student hands-on science scale but not with the teacher scale. With NELS, we found a relationship between the 8th and 10^{th} grade teacher scale and test scores but little relationship when using the student scale in 10^{th} grade. No relationship was found when using NELS 12^{th} grade data, teacher and student.

However, our findings in regard to Hypothesis 1 need to be further evaluated. We must consider the quality of the data and its appropriateness before concluding that the mixed findings do not allow us to make a conclusion regarding Hypothesis 1.

Evaluating the RAND and NELS:88 Results Based on the Strengths and Weaknesses of the Data

The RAND and NELS data sets differ in their strengths and weaknesses and this is what makes them complimentary. When considering the inconsistent evidence provided by the results from both data sets we wish to place greater emphasis on the results derived from the strengths of each. Here we examine the strength of the data sets in regards to three issues: 1) the teacher scale results, 2) the student scale results, and 3) the NELS 12th grade results.

Regarding the teacher scale, NELS has the better data on teacher reports of hands-on science. The RAND data contains a small sample of teachers with little variation in their reports of hands-on science. For this reason, it is not unexpected that we would not find a relationship between the teacher scale and test scores when using the RAND data. NELS has a large teacher sample with greater variation in teacher reports. If we had failed to find the positive relationship using the NELS, it would be strong evidence that the relationship does not exist. However, we did find such a relationship when using NELS (8th and 10th grade data) and we can attribute this, in part, to the larger teacher sample available. For this reason, we have greater confidence in our results for the teacher scale when using the NELS data than the RAND data. Furthermore, our confidence is greater in the NELS results because of the additional covariates that have been controlled for.

For the student scale, the RAND data is stronger. RAND's student survey asked questions well related to the level of hands-on science in the classroom including key ones on how often experiments were done and scientific materials were used in class. NELS did not ask students in 8th grade about the level of hands-on science so no comparison with the RAND data can be done for the 8th grade. In 10th grade students were asked several questions but the key one on how often experiments were done was not included. Therefore the 10th grade student scale is lacking a crucial item and we are not surprised that it shows little relationship with test score. For this reason, we have greater confidence in our results for the student scale when using the RAND data.

Both the 12th grade student and teacher scales showed no relationship with test score. The teacher scale was similar in construction to the scales of the earlier grades and the student scales contains the crucial item on how often experiments were done. Therefore the quality of the items making up the scale does not reduce our confidence in the results.

However, there are two other concerns regarding the 12th grade data. The first is the reliability of the 12th grade (and 10th grade) student scale. As noted in Chapter 5, our descriptive analysis found opposite results for the 10th grade versus 12th grade student scale when students were broken down into racial/ethnic and student ability groups. Specifically, Black reported the lowest scale in 10th grade but highest in 12th. The student scale also flipped from high to low in 10th versus 12th grade for the highest ability rank and class achievement. This concern over the student scale reduces our confidence in the 12th grade results based upon it.

Second, 12th grade students are very different from 8th grade students. Abstract thinking should be further developed by 12th graders and the concrete illustrations of hands-on science may be less beneficial for their learning. For this reason, they may not respond to hands-on science the same way that 8th graders do and we would expect the relationship of hands-on science to test score to be weaker for older students. For this reason, we have less confidence that the results using the 12th grade data make a proper comparison with those of 8th grade.

Our evaluation leads us to differentiate among the results providing us with more confidence in those based upon the RAND 8th grade student scale and the NELS 8th and 10th grade teacher scales. From these results, we argue that hands-on science is positively related to test scores. The evidence is stronger for a relationship with multiple choice test scores as it comes from two separate surveys using different multiple choice tests. The results from the NELS data makes us more confident that the relationship between hands-on science and multiple choice test scores does indeed remain when a fuller array of covariates are controlled.

The magnitude of the relationship was found to be substantial for both data sets. An increase of 1 point in the 5 point hands-on scale leads to an increase of almost 0.2 of a standard deviation of the test score for the multiple choice test in the RAND analysis and 0.1 in the NELS 8th grade analysis. If we consider a shift from the lowest to the highest level of hands-on science (from 1 to 5 points), the coefficients would convert to four times their value, a magnitude equivalent (though opposite in the sign) to what we saw for the negative relationships of Hispanic or Black to test score.

In conclusion, our analysis provides three important findings. First, hands-on science is positively related to test scores for upper middle and lower high school students. Second, this relationship persists after controlling for additional variables that are closely related to science achievement and after carrying out a series of sensitivity analysis. Third, the relationship is substantial in that a full implementation of hands-on science is equivalent to the negative association of minority status (Black and Hispanic) with test score. Not enough evidence was found in support of the two other hypotheses that hands-on science would be more strongly associated with performance test scores versus multiple choice test scores or that its association with test score would depend upon student ability rank. Further research, discussed below, could further confirm these findings and address the weaknesses of our data.

Policy Implications

After using the strengths of the data to evaluate the results, we find a positive relationship between hands-on science and test score. This finding supports a continued emphasis on the promotion of hands-on science that began in the 1960s, with an exception for upper high school. The finding of a relationship with both multiple choice and performance test scores should make this promotion more attractive. In those states or districts that intend to continue to rely on multiple choice tests, the use of hands-on science can support efforts to increase scores without the fear that the increased time required for hands-on science will harm student test scores. States or districts that adopt performance assessments often promote increase hands-on science under an assumption that the two are linked. Our results provide evidence to confirm this assumption. States and districts using a combination of test types may have been torn in the approach to take toward increasing scores. Our results show that an emphasis on hands-on science can support gains in both types of test scores.

A continued emphasis on promoting hands-on science will require greater attention to the practical constraints to its use (discussed in Chapter 2) including the need for logistical support and adequate training whose lack helped block the adoption of the curricula developed in the 1960s (Chapter 3). It is clearer today that supporting the teaching of hands-on science requires more than just developing new curricula. Others issues need to be addressed including the recurrent needs in training faculty and providing materials in a timely and affordable basis, addressing faculty turnover, offsetting deficits in faculty content knowledge and classroom management skills, and ensuring long-term administration support for this work. In addition, the requirement for greater time per topic under hands-on science (discussed in Chapter 2) must be taken into account. This can be partly addressed by reducing the repetitive busy work aspect of hands-on science that has been criticized throughout its history. But, consideration will also have to be given to a reduction in content coverage both in the curriculum and in the standardized tests.

The finding of a positive relationship does not support the current emphasis of science reform to temper the use of hands-on science. Instead, it suggests that further research should first be done on the instructional methods proposed by reform to determine if there is evidence for their reduced emphasis on hands-on science.

Additionally, current efforts to focus on inquiry as the primary instructional approach for hands-on science need to be reconsidered. We were not able to separate out the relationship by instructional approach so our results are based on the average effects of the various instructional approaches. As inquiry has only been promoted for a relatively short period of time it is unlikely to be the predominant instructional approach used with hands-on science and so is unlikely to be the primary source of the relationship of hands-on science and test scores. Further research should be done on all available instructional approaches before we discard some of them that may be contributing to positive student outcomes.

The finding that the positive relationship of hands-on science and test score does not differ by type of test has implications for testing programs. If multiple choice tests capture the benefits of hands-on science as well as performance tests, then the promotion of the increased used of performance tests cannot be justified on these grounds. An alternate view of our results could be that both the performance and multiple choice tests used in this analysis were poor at capturing process skills. For this reason, further research using other versions of both types of tests should be done in an attempt to extend our analysis.

Our inconsistent findings on whether ability rank affects the link of hands-on science to achievement should reduce criticisms of its widespread application among all students. That the evidence does not clearly show an adverse relationship for high ability students should reduce resistance from academic oriented students and their parents who fear that hands-on science might reduce the breadth of knowledge needed to succeed in standardized tests.

In considering these policy implications, several caveats should be kept in mind. First, our conclusion of a positive relationship of hands-on science and test score is based on our evaluation of the mixed results using the strengths of the data sets. Others might evaluate the results differently and give greater weight to our findings that did not show such a relationship. Second, our conclusion concerns a positive relationship but does not prove a causal relationship between hands-on science and test score. If some other non-controlled factor is responsible for both greater levers of hands-on science and higher tests, then policies to promote hands-on science may not lead to higher test scores unless they also raise levels of the unknown factor. Third, our analysis was not able to control for quality of instruction nor instructional approach (the latter was discussed above). However, our inability to consider quality of hands-on instruction leads to conservative results. By definition, higher quality instruction leads to greater student learning and higher test scores. We would then expect to find a stronger relationship between high quality hands-on science and test score than the relationship we found which was based on an unknown average level of quality.

In sum, our results provide evidence in favor of the above discussed policy options. Additional research will be necessary to confirm our findings, further address some of the inconsistencies we found as well as overcome the caveats we had to place upon our policy options due to research issues we could not resolve.

Further Research

Further research needs identified in this work can be broken down into two categories: research to confirm and extend our findings and research to support the better use of hands-on science in the classroom.

Because the two data sets were in many ways complementary rather than confirmatory, there are several findings whose replication would provide additional evidence to support the policies described above. The results from the NELS analysis would be strengthened using similar data that also contained performance test scores. A data set containing both more covariates and performance test scores could also be used to confirm the lack of a differential relationship by type of test. Furthermore, the collection of this type of data longitudinally that also contained consistent items on the level of hands-on science across grades would allow a stronger test of the relationship of hands-on science to test score including some consideration of causality.

We examined two multiple choice tests from separate respected sources and two performance tests developed by RAND. Similar results using different standardized tests would strengthen confidence in our results. The policy impact of this further research would be greater if the tests chosen were already in wide use.

Confirmation of a more basic issue would be to address the validity of hands-on scales obtained from survey data using such techniques as classroom observation or teacher and student logs. Surveys on hands-on science to both students and teachers could be given and compared to results from these techniques. These comparisons could be used to improve survey items and help develop items to identify quality and instructional approach. This work might also help explain inconsistencies in results from teacher versus student reports on the level of hands-on science.

Research for the better application of hands-on science is primarily concerned with the instructional approaches used in providing hands-on science. The current choices in the approaches are primarily made on theoretical grounds or convenience. Specific studies on the approaches used in the classroom and the relationship with test scores could be done to determine the best approach or mix to use. The inclusion of items in the national surveys on what instructional approaches are used with hands-on science would be a step forward in this direction.

In addition, further analysis on upper high school grades could examine the uniqueness of upper high school science instruction and learning. This could help determine empirically if there is a difference in the relationship of hands-on science and test score between middle school and upper high school students. The lack of evidence from our work to support the use of hands-on science in upper high school requires further research to determine its role there.

References

American Association for the Advancement of Science, "AAAS Project 2062: Middle Grades Science Textbooks Evaluation: Criteria for Evaluating the Quality of Instructional Support," AAAS website, 1999.

American Association for the Advancement of Science, "Project 2061: Update 1997", Washington, DC: AAAS, 1997.

Anderson, Lorin, Ryan, Doris and Shapiro, Bernard, The IEA Classroom Environmental Study: Interantional Studies in Education Achievement, Volume 4, New York: Pergamon Press, 1989.

Arons, Arnold, "Achieving Wider Scientific Literacy," Daedalus 112(2):91-122, 1983.

Atkinson, Elaine, "Learning Scientific Knowledge in the Student Laboratory" in Elizabeth Hegarty-Hazel (ed.), The Student Laboratory and the Science Curriculum, New York: Rutledge, 1990.

Baxter, Gail, Shavelson, Richard, Goldman, Susan and Pine, Jerry, "Evaluation of Procedure-Based Scoring for Hands-on Science Assessment," Journal of Education Measurement 29(1):1-17, Spring 1992.

Black, N., "Better Demonstration in Physics", School Science and Mathematics 30:366-373, 1930.

Blumberg, Fran, Epstein, Marion, MacDonald, Walter and Mullis, Ina, "A Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics: Final Report – Part I," National Assessment of Educational Progress, November 1986.

Breddeman, Ted, "Effects of Activity-Based Elementary Science on Student Outcomes: A Quantitative Synthesis," Review of Educational Research 53(4):499-518, Winter 1983.

Bruner, Jerome, The Process of Education, New York: Vintage, 1960.

Burkham, David, Lee, Valerie, and Smerdon, Becky, "Gender and Science Learning Early in High School: Subject Matter and Laboratory Experiences," American Educational Research Journal, 34(2):297-331, Summer 1997.

Burnstein, Leigh, McDonnell, Lorraine, Van Winkle, Jeannette, Ormseth, Tor, Mirocha, Jim and Guiton, Gretchen, "Validating National Curriculum Indicators", Rand Corporation, Santa Monica, CA, 1995.

Carpenter, W., Certain phases of the Administration of High-School Chemistry,
New York: Teachers College, Columbia University, 1925.

Champagne Audrey, Klopfer, Leopold and Gunstone, Richard, "Cognitive Research and the Design of Science Instruction," Educational Psychologist 17(1):31-53, 1982.

Comber, L. and Keeves, J., Science Education in 19 Countries: An Empirical Study, New York: Halsted Press, 1973.

Cronbach, Lee and Richard Snow, Aptitude and Instructional Methods, New York: Irvington Publishers, 1981.

Cunningham, H., "Lecture Method versus Individual Laboratory Method in Science Training: A Summary," Science Education 30:70-82, 1946.

Deboer, George, A History of Ideas in Science Education, New York: Teachers College Press, 1991.

Doran, Rodney and Tamir, Pinchas, "Results of Practical Skills Testing," Studies in Educational Evaluation 18:365-392, 1992.

Driver, R., "Pupils Alternative Frameworks in Science," European Journal of Science Education 18:365-392, 1992.

Driver, R. and Bell, B, "Students Thinking and the Learning of Science: A Constructivist View," School Science Review 67(240):443-456, 1986.

Dunbar, Stephen, Koretz, Daniel, and Hoover, H., "Quality Control in the Development and Use of Performance Assessments," Applied Measurement in Education 4(4):289-303, 1991.

Eylon, Bat-Sheva and Linn, Marcia, "Learning and Instruction: An Examination of 4 Research Perspectives in Science Education", Review of Educational Research 58(3):251-301, Fall 1988.

Fetters, William, Stowe, Peter, and Owings, Jeffrey, "High School and Beyond: A National Longitudinal Study for the 1980's: Quality of Responses of High School Students to Questionnaire Items," U.S. Department of Education, NCES, Washington, DC, September 1984.

Friedler, Yael and Tamir, Pinchas, "Life in Science Classrooms at Secondary Level," in Elizabeth Hegarty-Hazel (ed.), The Student Laboratory and the Science Curriculum, New York: Rutledge, 1990.

Gage, N. and Berliner, D., Educational Psychology, Boston: Houghton Mifflin Co., 1984.

Glynn, Shawn and Duit, Reinders (eds.), Learning Science in the Schools, Mahwah, NJ: Lawrence Earlbaum Associates, 1995.

Hamilton, Laura, Nussbaum, Michael, Kupermintz, Haggai, Kerkhoven, Joannes and Snow, Richard, "Enhancing the Validity and Usefulness of Large-Scael Educational Assessments: II. NELS:88 Science Achievement", American Educational Research Journal 32(3):555-581, Fall 1995.

Harmon, M. and Mungal, C., "The Influence of Testing on Teaching Math and Science in Grades 4-12: Appendix C: An Analysis of Standardized and Text-Embedded Tests in Science," Chestnut Hill, MA: Center for the Study of Testing, Evaluation and Educational Policy, 1992.

Herman, J., Aschbacher, P. and Winters, L., A Practical Guide to Alternative Assessments, Alexandria, VA: Association for Supervision and Curriculum Development, 1992.

Hodson, Derek, "Laboratory Work as Scientific Method: Three Decades of Confusion and Distortion," Journal of Curriculum Studies 28(2):115-135, 1996.

Hoffer, Thomas and Moore, Whitney, "High School Seniors' Instructional Experiences in Science and Mathematics," NCES 95-278, U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, Washington, D.C.: U.S. Government Printing Office, February 1996.

Hoffer, Thomas, Rasinski, Kenneth, and Moore, Whitney, "Social Background Differences in High School Mathematics and Science Coursetaking and Achievement", Report #95-206, National Center for Education Statistics, U.S. Department of Education, 1995.

Hofstein, Avi and Lunetta, Vincent, "The Role of the Laboratory in Science Teaching: Neglected Aspects of Research,""Review of Educational Research 52(2):201-217, 1982.

Hofstein, A. and Yager, R., "Societal Issues as Organizers for Science Education in the 80's", School Science and Mathematics 82:539-547, 1982.

Horn, Laura, Hafner, Anne, and Owings, Jeffrey, "A Profile of American Eighth-Grade Mathematics and Science Instruction: NELS:88." NCES 92-486, U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, Washington, D.C.: U.S. Government Printing Office, June 1992.

Hurd, Paul, New Directions in Teaching Secondary School Science, Chicago: Rand McNally, 1970.

Huxley, T., Science and Education, New York: Appleton, 1899.

Jones, Lee, Mullis, Ina, Raizen, Senta, Weiss, Iris, and Weston, Elizabeth, The 1990 Science Report Card: NAEP's Assessment of Fourth, Eighth and Twelfth Graders, Princeton, NJ: ETS, March 1992.

Karpus, Robert, "Science Teaching and the Development of Reasoning," Journal of Research in Science Teaching, 14:169-175.

Kelly, P. and Lister, R., "Assessing Practical Ability in Nuffield A-Level Biology," in J. Eggleston and C. Newbould (eds.) Studies in Assessment, London: English University Press, 1969.

Klopfer, Leopold, "Learning Scientific Enquiry in the Student Laboratory", in Elizabeth Hegarty-Hazel (ed.) The Student Laboratory and the Science Curriculum, New York: Rutledge, 1990.

Kojima, Shugeo, "IEA Science Study in Japan with Special Reference to the Practical Test," Comparative Educational Review pp.262-267, June 1974.

Koran, Mary Lou and John Koran, "Aptitude-Treatment Interaction Research in Science Education," Journal of Research in Science Teaching 21(8):793-808, 1984.

Lapoint, Archie, Mead, Nancy and Phillips, Gary, A World of Differences, ETS Report #19-CAEP-01, Princeton, NJ: January 1989.

Lapointe, Archie, Askew, Janice and Mead, Nancy, Learning Science, ETS Report #22-CAEP-02, Princeton, NJ: February 1992.

Lawson, Anton, Science Teaching and the Development of Thinking, Belmont, CA: Wadsworth Publishing, 1995.

Lee, Valerie and Burkam, David, "Gender Differences in Middle Grade Science Achievement: Subject Domain, Ability Level and Course Emphasis" Science Education 80(6):613-650, 1996

Lowry, Lawrence, "The Scientific Thinking Process", Berkeley, Full Option Science System, Lawrence Hall of Science, University of California, 1992.

Madaus, G., West, M., Harmon, M., Lomax, R., and Viator, K., "The Influence of Testing on Teaching Math and Science in Grades 4-12," Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation and Educational Policy, 1992.

Madigan, Timothy, "Science Proficiency and Course Taking in High School", Report # 97-838, National Center for Education Statistics, U.S. Department of Education, April 1997.

Matthews, Michael, Science Teaching: The Role of History and Philosphy of Science, New York: Rutledge, 1994.

Mayer, Daniel, "Measuring Instructional Practice: Can Policymakers Trust Survey Data?", Educational Evaluation and Policy Analysis 21(1):29-45, Spring 1999.

Mehrens, William, "Using Performance Assessment for Accountability Purposes," Educational Measurement: Issues and Practice 11(1):3-20, Spring 1992.

Mervis, Jefrey, "Mixed Grades for NSF's Bold Reform of Statewide Education," Science Vol. 282:1800-1805, Dec. 4, 1998.

Miller, M. and Legy, S., "Alternative Assessment in a High Stakes Environment," Educational Measurement: Issues and Practice 12(2):9-15, Summer 1993.

Mullis, Ina and Jenkins, Lynn, The Science Report Card: Elements of Risk & Recovery Trends and Achievements Based on the 1986 National Assessment of Educational Progress, ETS Report #17-S-01, Princeton, NJ: September 1988.

Murnanne, Richard and Raizen, Senta (eds.), Improving Indicators of the Quality of Science and Math Education in Grades K-12, Washington, D.C.: National Academy Press, 1988.

NCES, OERI, U.S. Department of Education. 1994. "Second Follow-Up: Student Component Data File User's Manual: NELS:88," NCES:94-374, Washington, DC.

NCES, OERI, U.S. Department of Education. 1995. "Psychometric Report for the NELS:88 Base Year Through Second Follow-up", Washington, D.C.

National Education Association, Reorganization of Science in Secondary Schools: A Report of the Commission on the Reorganization of Secondary Education, U.S. Bureau of Education, Bulletin 35, Washington, DC: U.S. Government Printing Office, 1920.

National Research Council, National Science Education Standards, Washington, DC: National Academy Press, 1996.

National Science Teachers Association, "NSTA Position Statement on School Science Education," The Science Teacher 38:46-51, 1971.

National Society for the Study of Education, A Program for Teaching Science: Thirty-First Yearbook of the NSSE, Chicago: University of Chicago Press, 1932.

National Society for the Study of Education, Science Education in American Schools: Forty-Sixth Yearbook of the NSSE, Chicago: University of Chicago Press, 1947.

Nussbaum, Michael, Hamilton, Laura and Snow, Richard, "Enhancing the Validity and Usefulness of Large Scale Educational Assessments: IV. NELS:88 Science Acheivement to 12th Grade", American Educational Research Journal 34(1):151-173, Spring 1997.

Office of Technology Assessment, U.S. Congress, Testing in American Schools: Asking the Right Questions, Washington, D.C.: U.S. GPO, 1992.

Peng, Samuel and Susan Hill, "Understanding Racial-Ethnic Differences In Secondary School Science and Mathematics Achievement," NCES 95-710, Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, February 1995.

Piaget, Jean, To Understand is to Invent: The Future of Education, New York: Grossman, 1973.

Porter, Andrew, "Developing Opportunity-to-Learn Indicators of the Content of Instruction: Progress Report," Wisconsin Center for Education Research, School of Education, University of Wisconsin, Madison, September 1995.

President's Scientific Research Board, Science and Public Policy, Washington, DC: U.S. Government Printing Office.

Resnick, Lauren, and Klopfer, Leopold, "Toward the Thinking Curriculum: An Overview," in Lauren Resnick and Leopold Klopfer (eds.) Toward the Thinking Curriculum: Current Cognitive Research, 1989 Yearbook of the Association for Supervision and Curriculum Development, ASCD #610-89012, ASCD 1989.

Resnick, Lauren and Resnick, Daniel, "Assessing the Thinking Curriculum: New Tools for Educational Reform," in Bernard Gifford and Mary O'Connor (eds.) Future Assessments: Changing Views of Aptitude, Achievement and Instruction, Boston: Kluwer Academic Publishers, 1992.

Rossi, P., Wright, J. and Anderson, A. (eds.), Handbook of Survey Research, Academic Press, San Diego, 1983.

Schwab, Joseph, The Teaching of Science, Cambridge, MA: Harvard University Press, 1962.

Shavelson, Richard, Baxter, Gail, and Gao, Xiahong, "Sampling Variabilty of Performance Assessments", Journal of Educational Measurement 30(3):215-232, Fall 1993.

Shulam, Lee and Tamir, Pinchas, "Research on Teaching in the Natural Sciences," in Robert Travers (ed), Second Handbook of Research on Teaching, Chicago: Rand McNally, 1973.

Shymansky, James, Kyle, William, Alport, Jennifer, "The Effects of Science Curricula on Student Performance, Journal of Research in Science Teaching 20(5):387-404, 1983.

Smith, A. and Hall, E., The Teaching of Chemistry and Physics in the Secondary School, New York: Longmans, 1902.

Spencer, H., Education: Intellectual, Moral and Physical, New York: Appleton, 1864.

Stecher, B. and Klein, S., (eds.), "Performance Assessments in Science: Hands-on Tasks and Scoring Guides" (MR-660-NSF), Santa Monica, CA: RAND, 1996.

Sudman, Seymour and Bradburn, Norman, Asking Questions: A Practical Guide to Questionnaire Design, Jossey-Bass, San Francisco, 1991.

Sullivan, Christine and Weiss, Andrew, "Student Work and Teacher Practices in Science," NCES 1999-455, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, July 1999.

Tamir, Pinchas, "The Practical Mode – A Distinct Mode of Performance in Biology," Journal of Biological Education 6:175-182, 1972.

Tamir, Pinchas and Doran, Rodney, "Science Process Skills in 6 Countries," Second IEA Science Study, mimeo, no date.

U.S. Office of Education, Life Adjustment for Every Youth, Washington, DC: U.S. Government Printing Office, 1951.

U.S. Office of Education, Education for the Talented in Mathematics and Science, Washington, DC: U.S. Government Printing Office, 1953.

Valliga, Michael, "The Accuracy of Self-Reported High School Course and Grade Information," ACT Research Report Series 87-1, American College Testing Program, Iowa City, November 1986.

Weiss, Iris, "1977 National Survey of Science, Mathematics and Social Studies Education Highlights Report," in The Status of Pre-College Science, Mathematics and Social Studies Educational Practices in U.S. Schools: An Overview and Summaries of Three Studies, Washington, DC: U.S. Government Printing Office, 1978.

White, Richard and Tisher, Richard, "Research on Natural Sciences", in Merlin Wittrock (ed.) Handbook of Research on Teaching, chapter 30, NY:Macmillan, 1985.

Wiggens, Grant. "A True Test: Towards More Authentic and Equitable Assessment", Phi Delta Kappan 703-713, May 1989.

Wellington, Jerry, "Practical Work in Science: Time for a Re-appraisal," in Jerry Wellington (ed.) Practical Work in School Science, New York: Rutledge, 1998.

Williams, L., "The Urban Systemic Initiatives (USI) Program of the National Science Foundation: Summary Updates," Washington, D.C.: NSF, 1998.

Winship, Christopher and Radbill, Larry, November 1994, "Sampling Weights and Regression Analysis," Sociological Methods & Research 23(2):230-257.

Yaeger, Robert, Engen, Harold and Snider, Bill, "Effects of the Laboratory and Demonstration Methods Upon the Outcome of Instruction in Secondary Biology," Journal of Research in Science Teaching 6:76-86, 1969.

Youmans, E. (ed.) The Culture Demanded by Modern Life, New York: Appleton, 1867.